GEOSPATIAL RESEARCH INSTITUTE
LAND INFORMATION NEW ZEALAND

Towards Real-time Federation of Spatial Data for Hazard Risk Management

Report for the Cooperative Research Centre for Spatial Information (project phase 1)

Jeremy SIAO HIM FA[1], Kasey OOMEN[2], Jiakai LI[3], Rob DEAKIN[2], Ivana IVÁNOVÁ[1], David A. MCMEEKIN[1] and Matthew D. WILSON[3]

1. School of Earth and Planetary Sciences, Curtin University, Perth, Australia
2. Land Information New Zealand, Wellington, New Zealand
3. Geospatial Research Institute, University of Canterbury, Christchurch, New Zealand

CRC-SI project agreement: 3.24

Project Leaders:

Prof. Matthew WILSON[3]
Mr. Robert DEAKIN[2]
Dr. David A MCMEEKIN[1]

January 23, 2019
(final version)

Contents

Executive Summary

Hazard risk management requires the complex analysis of dynamic datasets. However, on a national scale, there are issues related to heterogeneity among datasets, data quality and data accuracy, which often need to be reconciled manually or with hard-coded rules. Much time and effort therefore has to be given to the reconciliation of the datasets that could otherwise be used for the modelling and prediction of hazards. To improve this balance, more automated integration processes are required. This would allow experts to focus on assessing the probable size and extent of hazards and the reduction of their impacts ahead of time.

In addition, because hazard datasets change regularly (e.g. rainfall and landscape for flood), the datasets need to be integrated in real-time whenever possible. In the spatial domain, automated real-time integration and analysis of data is most effectively achieved at a local scale; currently used Spatial Data Infrastructures (SDI) are not designed with automation in mind, but rather to minimise interoperability issues. As such, it is proposed that the current SDI be improved, shifting its capabilities towards those of a Spatial Knowledge Infrastructure (SKI), which would allow data to be integrated, and analyzed in real-time, enabling experts to refocus on the repercussion of hazards and the risks associated with them.

Dealing with natural hazards, whether planning or response, requires working with an increasing quantity of disparate spatial data. However, increasing data volumes are difficult to manage, particularly when data are distributed by multiple agencies. Consequently, from the hazard management perspective, data can be difficult to find and differing formats can make it challenging to produce consistent and integrated datasets across regional boundaries. This is particularly the case for data required with national coverage e.g. for a major hazard such as flooding which crosses agency boundaries. To aid with the development of flood risk mitigation, there is a need to increase data interoperability and accessibility, allowing improved mapping and assessment of flood hazard. This project aimed to address these issues through the specification of federated spatial data infrastructures, and the development of prototype APIs which leverage

these data into relevant information, building on Program 3 (Spatial Infrastructures) of the CRC-SI, which has explored federated data modelling methods to connect and provide a user generated view into different and disparate datasets. The project sought to develop outcomes from Program 3 towards the specifications of a workable pilot SKI implementation; to improve real-time access to hazard data and data federation. APIs were developed to illustrate the rapid production of information to flood managers, such as determination of the significance of measured or forecast rainfall.

A key Phase 1 objective was to engage with flood risk management practitioners (end users) to identify current practice and needs for data federation, identify data sources and types which are relevant for flood risk management, review existing spatial data infrastructure technologies and assess their capabilities for real-time data federation.

In Phase 2, to be implemented at a later date, the SDI identified in will be implemented alongside the toolbox developed. Together these will form the basis of a spatial knowledge infrastructure which allows for real-time federation of the flood risk data identified and is responsive to the needs of end users.

Through these two project phases, the core goal of the project is to enable the implementation of a SKI which comprises an SDI to host static and dynamic data of relevance to flood risk assessment, and APIs which leverage these data to produce information of relevance. Ultimately, the SKI system's uniqueness will be found in using spatial and temporal data from multiple data sources on-the-fly in situations where near real-time answers are required to real time, possibly life-threatening, situations.

Stakeholders play a crucial role in all parts of an SDI. Despite this, the implementation of SDIs has generally been restricted to enforcing policies and regulations on data collectors and data providers, often neglecting the consumers. A Spatial Knowledge Infrastructure (SKI) aims to step towards a next stage of SDI, where instead of giving data to consumers (and letting them extract the knowledge—the current approach), the knowledge is derived from within the SKI by automated real-time data integration and processing. The consumers then do not need to have specialised technical knowledge and can gain access

to knowledge on-demand without having to search for, understand, manipulate, and harmonise the data.

An SKI depends on ontologies where concepts are represented, inferred automatically, and queried over. This may be achieved using an entirely linked data approach (ideal, but which requires significant time, resource and funding commitments), an ontology-based data access approach (which limits some capabilities of linked data), or a hybrid approach, whereby the ontology-based approach can be used to address the gaps in linked data and be reduced or phased out as more links between data are created.

To understand the needs of flood risk practitioners towards the development of an SKI, two workshops in Wellington, New Zealand, and Christchurch, New Zealand were conducted in June 2018. The aims of the workshops were to identify the current methods used in flood risk management, the gaps in data and tools, the needs of the stakeholders, and whether there is demand for real-time analysis and integration of data.

The workshops were divided into three parts. The first part focused on datasets identification: which crucial datasets are available? what works well in terms of data provision? what are the current issues? and what improvements can be made to existing practice? The second part focused on identifying improvements to current tools and desired next generation tools. The third part focused on explaining the concept of an SKI, and to gather the users' thoughts on a potential SKI implementation.

The workshops identified the strong current demand, particularly from local government and utility operators, for flood modelling services. Typically these are to support local plans, flood protection scheme design and evaluation, and asset and infrastructure design and management. This demand is met by a mature and well skilled supplier community based within the private, research and local government sectors. In terms of outputs, we learned from participants that currently models tend to be created to simulate specific events (e.g. for a particular probability of occurrence / return period) rather than establishing a range of flood / depth probabilities at a particular location. Modelling to determine quantified risk (as a function of probability and consequence) is not common.

Key findings included a clear need for improvement in the current SDI; the lack of data consistency, quality, and real-time data feeds were identified as required improvements, alongside improved accessibility and documentation. The participants require data that is of high quality, consistent, easily accessible (e.g. centralised access), and of higher coverage.

Out of the three data categories assessed (source, pathway, receptor), receptor data requires most improvement; more data to be captured and made available.

A main tool identified to be a priority was a national flood model system, which would be consistent and provide live flood information. However, such a system should be open so that experts can input their own datasets, and extract their needed information. It would help in facilitating impact evaluation, enable the development of site specific models for infrastructure design activities, and allow near real-time scenario modelling.

It was also recognised that the flood models should be flexible enough to allow integration with existing risk assessment tools such as RiskScape.

The importance of understanding uncertainty in outputs, and supporting the visual interpretation of flood datasets was noted. The current inability to do these effectively were seen as key issues. It was agreed by the participants that more effort should be made to resolve them. Virtual reality and augmented reality were suggested as possible solutions.

The identified priorities for improved flood-risk modelling and the issues identified with the current system strongly support the implementation of a Spatial Knowledge Infrastructure (SKI). Approaches to improved data accessibility, consistency, interpretation, exploitation, customisation, and confidence can be tackled by taking an SKI-based approach. The participants saw benefits with the concept of an SKI for better flood-risk management. However, concerns were identified, particularly with regard to trust in the automation process, and in the ability to fund of such a system. Other significant issues that would require resolution include the governance of the SKI and the development of capability to deliver and maintain it.

To fully implement an SKI, the participation of its stakeholders is crucial. This project has taken this first step by engaging with flood modellers and managers to understand their needs and interests. Such users provide essential feedback to support system design and adoption. However, before a national scale implementation can occur, prototyping is required to ensure that the SKI meet these users' needs.

For the next phase of the project, use cases for the SKI will need to be identified alongside the requirements that the system need to fulfil. We suggest that stakeholder engagement workshops and presentations be held at regular intervals during the implementation of a flood SKI. This will result create and sustain demand for frequent updates and prototyping, which will drive a user-focused implementation of the SKI, resulting in an better likelihood of uptake.

1 Introduction

Hazard risk management requires the complex analysis of dynamic datasets that are often from different domains (e.g. land, water, meteorology, transport, geography, geology). To mitigate the risk of hazards, it is imperative that related data are processed in a timely and efficient manner allowing more time for preparation and response to disaster events.

Current methods to achieve this either rely on historic data to find patterns of re-occurrences, or use predictive models to determine the likelihood and potential damage of hazards. At a local scale, where the different datasets follow common standards and policies they are easier to integrate. On a national scale, there are issues related to heterogeneity among datasets (different syntaxes and schemas), data quality and data accuracy; these issues often need to be reconciled manually or with hard-coded rules. The majority of the labour involved is therefore focused on the reconciliation of the datasets instead of the modelling and prediction of hazards.

For a more proactive management of hazards, more automated integration processes are required to reconcile disparate and heterogeneous datasets. This, in return, would allow experts to focus on the extent of hazards and the reduction of casualties ahead of time. In addition, because hazard datasets change regularly (e.g. rainfall and landscape for flood), the datasets need to be integrated in real-time whenever possible.

In the spatial domain, automated real-time integration and analysis of data can be achieved but only at a local scale. The currently used Spatial Data Infrastructure (SDI) is not designed with automation in mind, but rather to minimise interoperability issues. As such, it is proposed that the current SDI be improved towards a Spatial Knowledge Infrastructure (SKI), which would allow data to be integrated, and analyzed in real-time, enabling experts to refocus on the repercussion of hazards and the risks associated with them. Duckham, Arnold, Armstrong, McMeekin, and Mottolini (2017) use the term Spatial Knowledge Infrastructure (SKI) to refer to an SDI that is semantically enabled allowing the real-time integration and analysis of data.

A shift to using more semantic enabled data would allow SDIs to infer new knowledge on-the-fly, and ease scalability and interoperability issues enabling more focused management of hazard risks.

1.1 Flood hazard

Flood is a significant hazard that can impair economic and social activities, cause damage to infrastructures, threaten lives, and have lasting geographical impacts with soil erosions and land slides (Ran & Nedovic-Budic, 2016). In 2017, from all natural disasters worldwide, flooding occurred 38% of the time, was responsible for 35% of total deaths, and impacted 60% of people experiencing hazards (EM-DAT, 2018). This hazard does not only affect the immediate flood plains but also secondary ones as the flood passes through them (Wrachien, Garrido, Mambretti, & Requena, 2012).

Flood hazards have traditionally been handled by attempting to control them using structures such as barricades or redirecting the flood. However, this approach requires constant maintenance and such 'flood prevention' methods can often be short-termed and insufficiently funded (Brown & Damery, 2002). For these reasons, there has been a shift towards flood-risk management, where the focus is not only on controlling the hazard but also on reducing its social impacts (Galloway, 2008). Flood-risk management is then a function of the likelihood of a hazard occurring, and the severity of its consequences (Ran & Nedovic-Budic, 2016).

However, the confidence in identifying a flood risk, and predicting its impacts is highly depended on the quality and accuracy of the data used—more accurate and better quality data increases the confidence of the prediction. Currently, the quality of core datasets relevant to flood risks is sub-optimal, or missing (e.g. national digital elevation models). In addition, flood-risk identification and prediction require the creation of flood models, which can vary depending on the methods used (Teng et al., 2017). Each of the different models and methods has their own benefits and disadvantages, but generally, the accuracy of a model is directly related to the intensity of the data processing, where more

accurate models require more processing power (Teng et al., 2017).

Further to the numerous modelling methods, flood-risk management requires a common governance of multiple disciplines (Hartmann & Driessen, 2017), which can lead to heterogeneities as the same concept can be defined differently (Wrachien et al., 2012). This issue is especially prevalent in hazard-risk management because the outputs of a user are required as the inputs for another user. Hence, the overall output of a hazard analysis is based on the interoperability of cross-domain datasets.

Common issues with current methods impairing flood-risk management have been identified from workshops and interviews lead by Land Information New Zealand (LINZ), University of Canterbury (UC), and Curtin University (CU). The main issues were that of data coverage, governance, policy, standardisation, accessibility, uncertainty, quality, and their real-time processing. As data coverage, governance, policy, and quality are issues to be addressed in the data collection phase, they are not discussed in this report. However, data standardisation, accessibility, uncertainty, and real-time processing are issues that can be resolved using technologies such as Linked Data (LD), cloud computing, and the Semantic Web in a next-generation Spatial Data Infrastructure referred to as a Spatial Knowledge Infrastructure (SKI).

1.2 Project overview

Dealing with natural hazards, whether planning or response, requires an increasing quantity of disparate spatial data. However, increasing data volumes are difficult to manage, particularly when data are distributed by multiple agencies. Consequently, from the hazard management perspective, data can be difficult to find and differing formats can make it challenging to produce consistent and integrated datasets across regional boundaries. This is particularly the case for nationally-relevant data for a major hazard such as flooding which crosses agency boundaries. To aid with the development of flood risk mitigation, there is a need to increase data interoperability and accessibility, allowing improved mapping and assessment of the flood hazard. This project aimed to address

these issues through the specification of federated spatial data infrastructures, and the development of prototype APIs which leverage these data into relevant information.

The project built on Program 3 (Spatial Infrastructures) of the CRC-SI, which has explored federated data modelling methods to connect and provide a user generated view into different and disparate datasets. In hazard risk and emergency management, time is a critical factor: SDI and SKI can help to remove barriers to access and provide linkages between necessary data, which are often held by different agencies at a national, regional or local level in both the public and private sectors, each of which maintain specific data for their own geographical areas and applications of interest. The project sought to develop outcomes from Program 3 towards the specifications of a workable pilot SKI implementation, to improve real-time access to hazard data. In this project, APIs were developed to illustrate the rapid production of information to flood managers, such as determination of the significance of measured or forecast rainfall. Project outputs will be used to validate and improve current methods of federating disparate datasets, as well as inform how 'semantic' data infrastructure can further be built.

1.3 Project aims and objectives

The broad aim of this project was to conduct work which will enable the development of a SKI for real-time data federation of relevant spatial and temporal data for agencies responsible for flood risk management and to develop prototype APIs which can leverage these data into products of relevance for flood risk mitigation. In Phase 1 of the project, reported here, the primary objectives were to:

1. Engage with flood risk management practitioners (end users) to identify current practice and needs for data federation;
2. Identify data sources and types which are relevant for flood risk management, including but not limited to infrastructure (e.g. stopbanks), topog-

raphy (e.g. LiDAR), temporal data (e.g. river gauge data), flood observational data, climatology, and existing flood zonation;

3. Review existing spatial data infrastructure technologies and assess their capabilities for real-time data federation; and

4. Develop a prototype spatial toolbox (API) which can allow rapid (real-time) visualisation of statistics of relevance to flood risk management (e.g. 100-year flood levels; rainfall intensity-duration-frequency curves), and provide an indicative map of likely flood areas.

In Phase 2, to be implemented at a later date, the SDI identified in (3) will be implemented alongside the toolbox developed in (4), which together will form the basis of a spatial knowledge infrastructure which allows for real-time federation of the flood risk data identified in (2) and is responsive to the needs identified in (1).

Through these two project phases, the core goal of the project is to enable the implementation of an SKI which comprises an SDI to host static and dynamic data of relevance to flood risk assessment, and APIs which leverage these data to produce information of relevance (Figure 1). These APIs will enrich data of relevance to flood risk management, through the provision of products such as the statistical characterisation and analysis of river flow or rainfall event data. From the outset, in order to achieve an SKI of most relevance, the project was guided by an assessment of needs from end users via workshops and interviews (Section 5); these will help ensure the inclusion of all relevant data and the appropriateness of the tools implemented.

1.4 Uniqueness and impact

Ultimately, the SKI system's uniqueness will be found in using spatial and temporal data from multiple data sources on the fly in situations where near real-time answers are required in possibly life-threatening situations. Once developed, the SKI will be an early demonstrable example of the ability to manage and derive

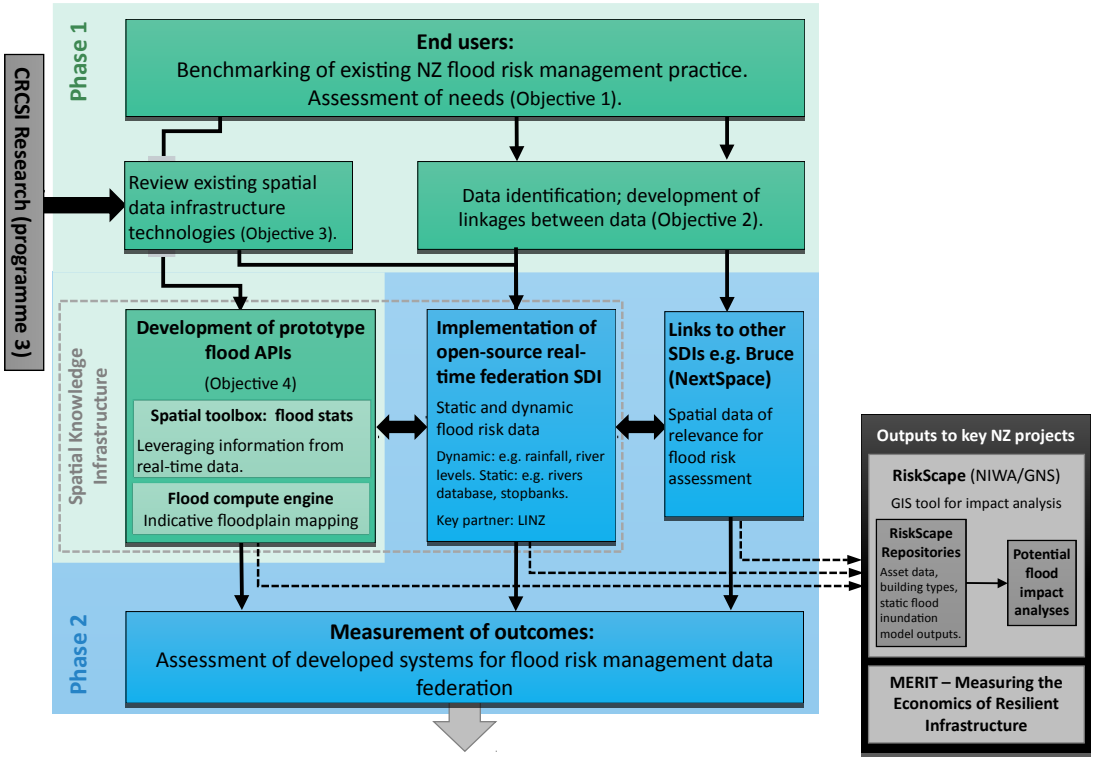

Figure 1: Project overview

new information from spatial and temporal data across multiple agencies at multiple levels of government for decision making in assessing and mitigating flood risk in a way which has not been possible previously.

The system developed will validate for hazard managers how a federated SDI can be used to develop or improve mitigation plans, through the provision of appropriate data and relevant information. The outputs of the project can help to improve the management of the significant risk of flooding posed within parts of New Zealand and Australia, particularly when mitigation is insufficient due to the lack of appropriate available data. Further, once the system has been demonstrated within the area of flood risk management and mitigation, it will be possible to develop it for use in other situations such as bush fire risk management and mitigation.

The project helps to inform New Zealand data suppliers of how better to make their information available, how their information is used and valued in flood management and how the private sector can leverage the project outputs to develop innovative products and services using the most up-to-date information. The system developed in project phase 2 will facilitate an increased up-take and usage of spatial data which exists in New Zealand.

1.5 Report scope and outline

The purpose of this report is to address the objectives of the project phase 1, as detailed in Section 1.3. The report also aims to provide comprehensive guidance towards the development of an SKI system, with a view towards future implementation in project phase 2. The scope of the report is restricted to flood risk management: while hazard risk is an ultimate end-goal, the development of systems for improved flood risk assessment and management are a challenging first step. The guidelines and framework presented in this report can be expanded to other hazards such as bush fires, earthquakes, and droughts. The technologies introduced in this report as part of an SKI are limited to the semantic Web, cloud computing, and Geospatial Information System (GIS) technologies.

Section 2, presents a review of SDI technologies with particular focus on

hazard-risk management, and proposes improvements to current SDI technologies towards more semantic-centred datasets. In Section 3, these ideas are developed into an outline of a prototype SKI, with a particular focus on flood risk management; in Section 4, a prototype API which may form part of a SKI is presented.

These ideas were assessed from the perspective of end-users or flood risk management professionals during two half-day workshops in June 2018. The results of these workshops are presented in Section 5 in three parts: assessing data needs (Section 5.1), assessing tool requirements (Section 5.3), and looking forward towards how we might develop and implement an SKI (Section 5.4).

Finally, in Section 6, the workshop outcomes are summarised and a pathway forward mapped, together with challenges faced in SKI implementation.

2 Next generation Spatial Data Infrastructures

There are many definitions for the concept of a Spatial Data Infrastructure (SDI) (GSDI Technical Working Group, 2009; Hendriks, Dessers, & van Hootegem, 2012). However, a commonality among all of them is the notion of an infrastructure (physical and virtual) to facilitate the use of spatial data. Infrastructure means a foundation that facilitates the spatial data supply chain from acquisition to delivery from different data sources (e.g. technologies, policies, and other institutional arrangements). The term 'use' can be expanded to include sharing of and access to spatial data. An SDI is therefore defined here as: the common technologies, policies, and other institutional arrangements for the acquisition, process, distribution, use, maintenance, and preservation of disparate spatial data to support its access, use, and sharing.

An SDI is required for disparate spatial data to interoperate, allowing for the retrieval of existing data and the creation of new solutions (e.g. the management of hazard-risks). However, this premise is based on either changing existing heterogeneous datasets to a new, homogeneous dataset, or to create the datasets anew. These cause problems because the growing number of spatial data means that a lot of data needs to be changed or re-surveyed—the latter might not be possible in certain instances (e.g. historical datasets)—which can be time and resource intensive. Further this requires businesses to cooperate and conform to the SDI, which might not necessarily benefit their current business models.

2.1 SDI in Flood-Risk Management

Major issues identified as part of the workshops and interviews were data coverage, governance, policy, standardisation, accessibility, uncertainty, quality, and their real-time processing. SDIs are meant to resolve the issue of standardisation and governance but they are often implemented regionally and for specific datasets. For the proper management of flood-risk, it was identified that a national SDI is needed so that the required cross-domain data are interoperable at large scale—flooding is often a cross-border hazard. This would mean a total

coverage of a nation, that is standardised and accessible.

The GSDI Technical Working Group (2009) identify the processes that are needed to truly resolve interoperability issues among datasets, they are:

1. Cross-border matching of datasets: this process can also be called data federation, it includes finding similar datasets that might be served by two independent service providers. Given that the syntax, structure, and semantics of the datasets can be different, it can be difficult to find similar datasets as no common language is used.

2. Cross-sector combinations of datasets from different repositories: this is data fusion where there is a need to create new data from existing data. There might be a case where a user query cannot be determined unless different datasets are fused together. For example, if no flood datasets exist in a particular region, they might have to be created by a combination of other datasets (e.g. rainfall, wind speed, elevation, and other geographic datasets).

3. Cross-type of data: this is syntactic heterogeneity where the datasets are represented in different data format (e.g. vector vs raster).

4. Overlap of the representation of the same entity that conflict: this is data conflation where the representation of the same entity from different data sources do not agree. The main issue with data conflation is finding the right datasets among a multiple conflicting ones, where in some cases, a compromise might have to be made to find a new common dataset that reconciles the conflicting ones.

Hence, it is commonly required to federate, fuse, and conflate interoperable datasets independently of their formats. This is often the case for flood-risk management because the datasets from different regions need to be gathered (federation), combined together to model the flood (fusion), and conflicting datasets need to be reconciled (conflation). Further, the data can be in different formats, requiring their seamless integration. For example, raster images from satellites and vectors from Web Feature Services.

2.2 SDI limitations

While a national SDI can resolve heterogeneities in datasets, the implementation of an SDI can be complex due to several reasons as mentioned by Hendriks et al. (2012). First, different SDIs can have different objectives at different granularity, from abstract to specific ones. Second, an SDI is heavily influenced by its stakeholders from a wide variety of disciplines, skill-sets, and view-points. Third, an SDI contains many aspects and components that often cannot be completely and/or logically listed. Fourth, an SDI is an organic concept that needs to evolve (technologically, politically, culturally, and socially) as the world changes. This evolution requires an SDI to be proactive to potential changes but at the same time reactive to unpredictable ones and these difficulties are magnified in large scale SDIs.

The current methods used to discover and access data using SDIs are often partially manual processes and can rely on human interactions and cooperation. For example, the use of geoportals which provide a repository for spatial data end-points require the manual registration of Web services (which need to be approved by a human user). In addition, they often only offer metadata keyword search and hence require domain specific knowledge (i.e. which keywords to search for), in order to efficiently find the required datasets. Further, geoportals are not cross-discoverable which limits their disoverability abilities.

When accessing data from an SDI, flood modellers still need to find, integrate, and process cross-domain datasets. However, these processes can be slow when done manually, or require bespoke customisation to automate processes. For better nationally accessible flood-management solutions, the traditional SDI needs to evolve to include automation, and machine processing which would allow data to be analysed and processed in real-time. For this to happen, there must be a standard for machines to comprehend the datasets and process them properly. A Spatial Knowledge Infrastructure (SKI) aims to address these limitations by adding semantics to existing SDIs enabling real-time data integration and processing.

2.3 Literature Review

Better hazard-risk management is an important research topic that is multi-faceted, containing different areas of research such as improving SDIs (Conti, Filho, Turra, & Amaral, 2018; Hendriks et al., 2012; Janowicz et al., 2010; Lutz, Sprado, Klien, Schubert, & Christ, 2009), automating the acquisition of data (Hu, Li, Lin, Chen, & Yang, 2018), mapping hazards (Kaur, Gupta, Parkash, & Thapa, 2018), analysing the hazards (Martínez-Graña et al., 2018; Metcalfe, Beven, Hankin, & Lamb, 2018; Sayers, Penning-Rowsell, & Horritt, 2018), and the prediction of their impact (Rauter & Winkler, 2018).

2.3.1 Current SDIs

Williamson (2003) describes the hierarchy of SDIs based on their coverage and scale; this is shown in figure 2. The SDIs at the lower levels have more detailed datasets, while the SDIs at a higher level deals with larger scale datasets. The corporate level SDI is the base level of the hierarchy (and hence has the most detailed datasets), and any SDI level above (local to global) is formed by integrating spatial datasets from the corporate SDIs to ensure their wider coverage. An SDI at a high level, therefore, depends on the datasets of SDIs at the lower levels.

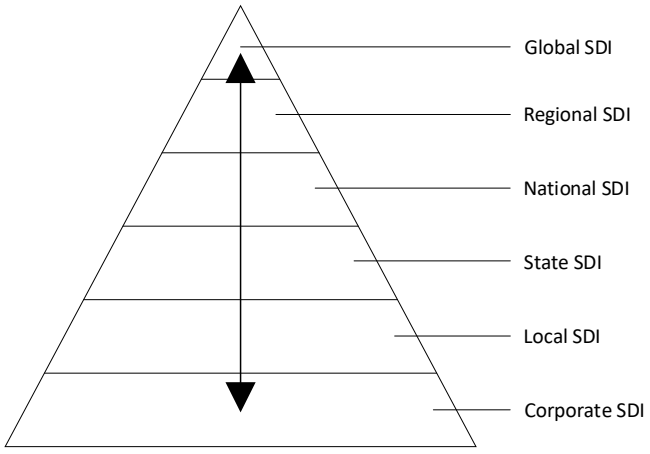

Figure 2: An SDI Hierarchy (Williamson, 2003, Chapter 2, p. 30)

Corporate SDIs

Corporate SDIs are concerned with the handling of spatial data at the corporation level. They ensure that datasets are stored and exposed appropriately to their stakeholders. SDIs at this level can also be of local and state level simultaneously, serving datasets at each of these scales.

Examples of corprate SDIs include: the Public Sector Mapping Agency (PSMA), Landgate, and the Department of Environment, Land, Water and Planning (DELWP) in Australia. These SDIs differ and have different functions. For instance, PSMA is a national aggregator in Australia and hence also part of the national level SDI. Landgate and DELWP cover datasets at the local and state level.

Local and State SDIs

Landgate and DELWP are examples of SDIs that cover datasets locally and at the state level. Landgate is in charge of the spatial data in Western Australia, and DELWP for Victoria. Each has their own standards and conventions that are subjective to their location, and the data they serve. For this reason, these datasets are not usually interoperable and require an SDI at a higher level to integrate them.

National SDIs

National SDIs can ensure that data from lower level SDIs are integrated to produce national datasets. For Australia, a central body such as PSMA is used to aggregate the disparate data. To integrate heterogeneous data, there are different methods that can be used; PSMA reconciles the heterogeneous data to their own local schema, but this causes duplication of data and data updates are slow. By contrast Geoscience Australia (GA) create national level datasets which may duplicate more detailed local data. In New Zealand, LINZ coordinate the collation of some local data in to national datasets (e.g. address, aerial imagery, elevation data) as well as producing national level datasets (e.g. road and river networks, building outlines), but again, more detailed and accurate data may be available for discrete local areas from local (sub-national) sources.

At an international level, the European Community has established the In-

frastructure for Spatial Information (INSPIRE). Its aim is to ease the sharing of spatial data across disparate public sector organisations in Europe and enable European level data integration (European Commission, 2018). The INSPIRE programme provides regulations and supporting guidelines regarding its SDI implementation that data provider must follow (European Commission, 2014). (The INSPIRE Directive has been transposed in to the laws and regulations of the Member States). This ensures a level of interoperability. While this avoids some of the issues experienced by PSMA and GA, some providers only provide data following the minimum rules and regulations, leading to a lower quality of data and a mixed levels of implementation across Member States.

Other initiatives in different countries in building a national SDI, include Zambia[1], New Zealand (New Zealand Government, 2011), Canada[2], and Australia[3]. These initiatives, however, have weaker mandates and rely on voluntary support, and are often therefore left either incomplete, outdated, or unimplemented.

Regional and Global SDIs

Two initiatives at the regional and global level include the Global SDI Working Group (GSDI-WG) (GSDI Technical Working Group, 2009) and the Global Earth Observation System of Systems (GEOSS)[4].

The GSDI-WG has been providing globally acceptable standards and practices to make spatial data interoperable at the global level; this information is available in their 'SDI Cookbook' (see GSDI Technical Working Group (2009)). They have acknowledged that their vision and mission has been adopted by internationally resourced organisations such as the United Nations, World Bank, and the Open Geospatial Consortium. As such, they have decided to cease their operations in 2018 due to having completed their mission of increasing the awareness

[1] http://www.nsdi.mlnrep.gov.zm/

[2] http://www.nrcan.gc.ca/earth-sciences/geomatics/canadas-spatial-data-infrastructure/10783

[3] http://ggim.un.org/knowledgebase/KnowledgebaseArticle50364.aspx

[4] https://www.earthobservations.org/geoss.php

of global interoperability (GSDI, 2018).

A different attempt for a global SDI is the Global Earth Observation System of Systems (GEOSS)[5] which allow multiple independent Earth observation, information and processing systems to coordinate, interact, and expose diverse information to a wide range of users. While GEOSS relies on countries to make their national data available following their specifications, there have been efforts by some countries to integrate their national data onto this globally accessible platform. For example, the European Global Earth Observation System of Systems[6] (EuroGEOSS). EuroGEOSS is a broker that aims for cross-discipline data integration globally using GEOSS. EuroGEOSS relies on already federated spatial data in the different domains. As such, before a solution such as EuroGEOSS can be implemented, the issue of heterogeneous data at the state level needs to be resolved first—Europe achieves this through INSPIRE. GEOSS, EuroGEOSS, and INSPIRE are examples where a global SDI (GEOSS) integrates SDIs at the regional and national level (EuroGEOSS), which in turn, uses SDIs at the state and local level (INSPIRE).

2.3.2 SDI Components for Better Hazard-Risk Management

Aside from the hierarchy of SDIs and their data integration, SDI components specific to this project's use case—hazard-risk management—must also be considered. The components reviewed in this section are the acquisition of data, the mapping and analysing the hazards, and the prediction of their impact.

Data Acquisition

The traditional method to obtain spatial data is through manual surveying of the earth, nowadays though, there are more advanced, precise, and automated ways to acquire spatial data. For example, Unmanned Aerial Vehicles (UAVs), satellites, sensors, autonomous vehicles, and radars. In hazard risk management,

[5] https://www.earthobservations.org/geoss.php

[6] https://ec.europa.eu/info/research-and-innovation/knowledge-publications-tools-and-data/knowledge-centres-and-data-portals/eurogeoss/about-eurogeoss_en

current, and more localised information is required to improve the management of risks to the individual. For such, recent works involve the creation of mobile applications to facilitate the reporting of incidents by common users (Olyazadeh, Sudmeier-Rieux, Jaboyedoff, Derron, & Devkota, 2017), and foraging social media platforms such as Twitter (Stowe et al., 2018).

However, a common challenge is the automated evaluation and selection of quality data. As data become more readily available online, human efforts to manually filter them are inefficient, and machines are required to quickly process the data instead. Hu et al. (2018) address this issue with the development of an ontology to semantically associate the observation capability of sensors. Their ontology is effective in helping sensor planners to make evidence-based sensor selection decision for the given flood observation task. Ontologies such as the one developed by Hu et al. (2018) can be used for properly selecting the right data from the varying sensors, improving the quality of the data obtained in real-time and reducing the amount of data needing processing.

Data Mapping

Further, machine automation is also needed to effectively process data in a timely manner. Due to the vast amount of data available, manually processing and filtering the data is inefficient. Neural networks[7] were used by Rauter and Winkler (2018) to automatically map new hazard zones. Rauter and Winkler (2018) focused on snow avalanche data, feeding their neural networks with historic data. Using this method, they managed to train their network using existing hazard maps and apply them to other regions, demonstrating that the knowledge gained from the network can be used in regions that do not have the hazard mapped. Using a similar idea, existing flood models can be used as training sets, and the existing known mappings can be applied to regions that are not mapped.

On a large scale, Vousdoukas et al. (2016) proposed a new methodology for mapping coastal flood hazard at European scale which accounts for the contri-

[7]Neural networks is a supervised machine learning technique, that uses interconnected processes to make statistical decisions based on pre-defined training sets.

bution of waves, uses improved inundation modelling and up-gradable physics-based framework. They found that the flood intensity index approach (Iw) (Dottori, Martina, & Figueiredo, 2016), and the LISFLOOD-FP (Bates & Roo, 2000) approach can be successfully used at large-scale.

Data Analysis and Prediction

Martínez-Graña et al. (2018) looked at the coastal vulnerability of the Menor Sea due to floods. The vulnerability was estimated using Remote Sensing techniques, and the risk of flooding was based on different time scenarios. Similar to Vousdoukas et al. (2016), they account for the contribution of waves in their proposal, and argue that their methodology can be used to accurately and effectively establish the sectors of high vulnerability (based on their physical characteristics and geographical position) to reduce the impact of flood innundation due to rises in sea level.

Metcalfe et al. (2018) propose a new simplified method to analyse the impacts of widely distributed enhanced hillslope storage on flood risk. They demonstrated that their simplified method can effectively model distributed hillslope storage within a less complex framework. On the other hand, Röthlisberger, Zischg, and Keiler (2017) found that spatial cluster analysis provided more information for prioritizing flood protection measures compared to the aggregation of data (which is more appropriate for nation-wide data analyses according to them).

However, hazards can also occur at the same time (e.g. an earthquake causing a tsunami). In this area, Kaur et al. (2018) investigated the application of current geospatial technologies for mapping multi-hazards, and the characterisation of their associated risk. They studied the region of Gangtok which is susceptible to multi-hazards such as earthquake, landslide, windstorm, flash floods and hailstorm. They concluded that the assessment of multi-hazard risks is more accurate at smaller scale, which can then give an effective spatial distribution of principal diverse risks at large scale.

It is evident that technologies used to produce higher quality and more ac-

curate hazard information are present, and continue to improve. Nonetheless, there is a gap where these technologies are not applied simultaneously to produce knowledge based on current hazard conditions. This is due to a lack of spatial infrastructure to bind these technologies, and issues with dataset interoperability and accuracy. The next section discusses the semantic Web technologies that can fill this gap. By incorporating semantic Web technologies to current SDIs, a knowledge infrastructure can be built which can facilitate decision making to better manage hazards—this knowledge infrastructure is referred to as a SKI in this report.

3 Developing a Spatial Knowledge Infrastructure for Flood Hazard Management

3.1 Overview

Stakeholders play a crucial role in all parts of an SDI. Despite this, the implementation of SDIs has generally been restricted to enforcing policies and regulations on certain groups: data collectors and data providers, often neglecting the consumers. Arnold (2016) describes this as a 'push' supply chain, where data get 'pushed' to the consumers based on anticipated demands rather than allowing the consumers to 'pull' the information they want.

A Spatial Knowledge Infrastructure (SKI) aims to step towards a next stage of SDI, where instead of giving data to consumers (and letting them extract the knowledge—the current approach), the knowledge is derived from within the SKI by automated real-time data integration and processing. The consumers then do not need to have specialised GIS knowledge and can gain access to knowledge on-demand without having to search for, understand, manipulate, and harmonise the data. This, in return, would inform the public about hazard risk, which could lead to more devolved adaption measures being taken and greater resilience to hazards (Hung, Lu, & Hung, 2018). In order to achieve this, technologies that allow the representation of knowledge are required. This can be achieved through the Semantic Web (Web 3.0).

3.2 The Semantic Web (Web 3.0)

Web 3.0 is envisioned, by the inventor of the Web, as an extension to the current Web where information is better defined hence allowing machines to 'comprehend' data (Berners-Lee, Hendler, & Lassila, 2001); in Web 3.0, machines can understand and process data, allowing for better automation for human consumption. The Resource Description Framework (RDF) (W3C, 2014a) was formulated to help standardise the semantic Web. RDF allows the representation of knowledge and is recommended for use by the World Wide Web Consortium

(W3C). It allows resources and their relationships to be expressed on the Web by using triples, which takes the form of <subject> <predicate> <object>, where the subject resource and the object resource are linked using a predicate resource. This creates data that are linked together, and is given the term Linked Data (LD).

Based on this framework, other semantic Web languages have emerged, such as (not all are listed here): Web Ontology Language (OWL) (W3C, 2004), which allows the expression of **what** a resource is on top of RDF which expresses **how** a resource is written; OWL2 (W3C, 2012), which provides more expressive terms and features to OWL; Shapes Constraint Language (SHACL) (W3C, 2017c) whose aim is to define the structure of an ontology for its validation; SPARQL Protocol and RDF Query Language (SPARQL—a recursive acronym) (W3C, 2013c), which enables the querying of triples; RDFa[8] (RDF in attributions) which allows users to annotate HTML elements using LD; and JSON-LD[9] which aims at facilitating the representation of LD *via* JavaScript Object Notation (JSON).

3.2.1 Ontology

The term ontology is used for the graph produced as a result of linking data together. Although the term ontology originates from philosophy—the study of existence and the nature of things —it has been adopted in computer science and is defined by Gruber (1993) as the explicit specification of domain knowledge, its concepts and their relationships; in this case, Linked Data.

Ontologies created by other people can be reused, reducing duplication of efforts and encouraging the sharing of concepts and knowledge. Upper ontologies are ontologies that have been abstracted to provide a starting point for more domain-specific ontologies; certain vocabularies and rules might have been expressed in the upper ontologies, which then do not need to be re-created but only reused. A few well-known upper ontologies are Friend of a Friend[10] (FOAF),

[8] https://rdfa.info
[9] http://json-ld.org/
[10] http://xmlns.com/foaf/spec/

Dublin Core[11], and Simple Knowledge Organization System[12] (SKOS).

Further, there are groups that create and publish more specific ontologies to aid ontology developers. One such group, founded by Google, Microsoft, Yahoo, and Yandex, can be found at `https://schema.org/`. It has a myriad of different ontologies that can be used by the community and is used by its founders. An example is in email reservations, where the reservation details are embedded using `schema.org` ontologies. The reservation details can then be automatically and seamlessly integrated by email assistant tools into calendars, reminders, notifications, and maps (Guha, Brickley, & Macbeth, 2016).

Other groups aim to promote other ontologies to help discover high-quality ones. For example, Linked Open Vocabularies (LOV)[13], which exposes vocabularies that adhere to certain criteria such as the stability of the URI, availability on the Web, use of standard formats, publication best practices, quality of the metadata and documentation, and versioning policy. There are also ontologies that are published by standard bodies such as W3C and OGC.

By using ontologies, concepts and knowledge can be shared across the Web, enabling non-expert user to utilise expert knowledge easily. In addition, RDF provides the standard ground for ontologies to be specified in, allowing machines to use this common language to find, and process data homogeneously. By adding this semantic component to existing SDI technologies, a shift to more automated and cross-domain applications is possible. In terms for flood-risk management, this means that flood inputs, processes, and outputs can all be uniformly computed by machines, allowing user to focus more on the impact of the flood rather than its simulation.

Further, with the aid of **reasoners**, the inference of new knowledge within ontologies can be achieved, where inferred relationships between concepts are automatically determined by the reasoner. This automatic inference of new knowledge based on existing data makes the Semantic Web a powerful machine-oriented platform for important decisions such as the management of risk haz-

[11] `http://dublincore.org/`
[12] `https://www.w3.org/2004/02/skos/`
[13] `http://lov.okfn.org/dataset/lov/about`

ards. The intermediary role of people from a traditional SDI is reduced, and a lot of the complex processes and information can be done by machines instead. For example, human users will no longer have to manually find the various data Web services available as all the data would be linked in a unified graph. Syntactic interoperability will not be an issue either, as all the data will be available in a Semantic Web language (which are interchangeable). Rules implemented by human users can be added in a local ontology to prevent conflicting insertion of data, and languages such as SHACL can be used to validate ontologies against a preferred structure.

3.3 SKI Architecture

Duckham et al. (2017, p. 4) define an SKI as: 'a network of data, analytics, expertise and policies that assist people, whether individually or in collaboration, to integrate in real time spatial knowledge into everyday decision-making and problem solving.'

Based on this definition, Arnold, McMeekin, Ivánová, and Armstrong (2018) extend an SDI to an SKI by adding two components:

1. Knowledge Representation; and
2. Analytics.

Knowledge can be represented using ontologies, and analytics are the processes that are enabled by proper knowledge representation. As such, a major extension to an SDI is the representation of data as ontologies instead of conventional documents.

As illustrated in figure 3, an SKI can be seen as a three-tiered architecture, with the resources being made available through the Web (bottom section), customised user applications interfacing with the SKI (top section), and the SKI broker (middle section) extending existing SDI components to enable the computation of complex processors (i.e. real-time federated reasoning, inferencing of new knowledge, and analysis of data).

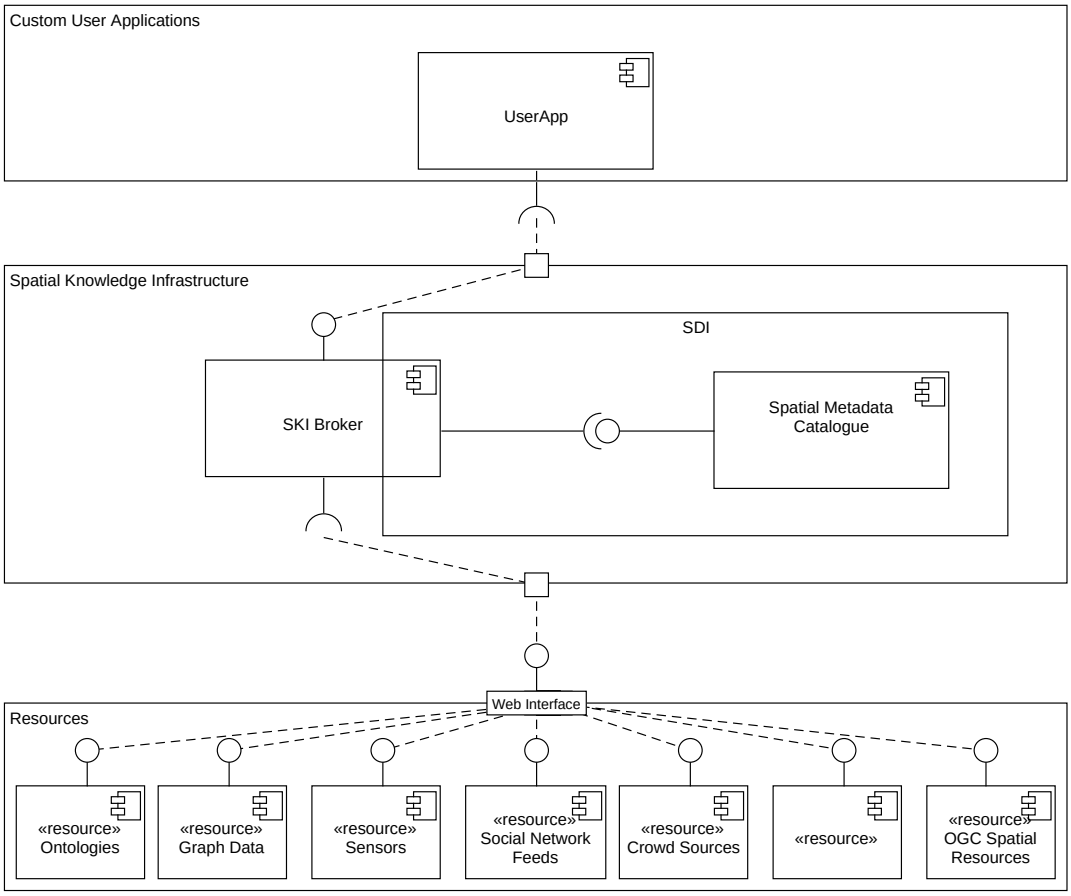

Figure 3: SKI Architecture

3.3.1 Components of an SKI Broker

While an SKI broker makes use of existing infrastructures as much as possible, there are components that are unique to it. For one, to optimally make use of ontologies, special databases called triple stores or graph-based databases must be used; these are referred to as Linked Data stores here. With RDF, rules can be inserted and the ontology reasoned over to find new relationships between data automatically.

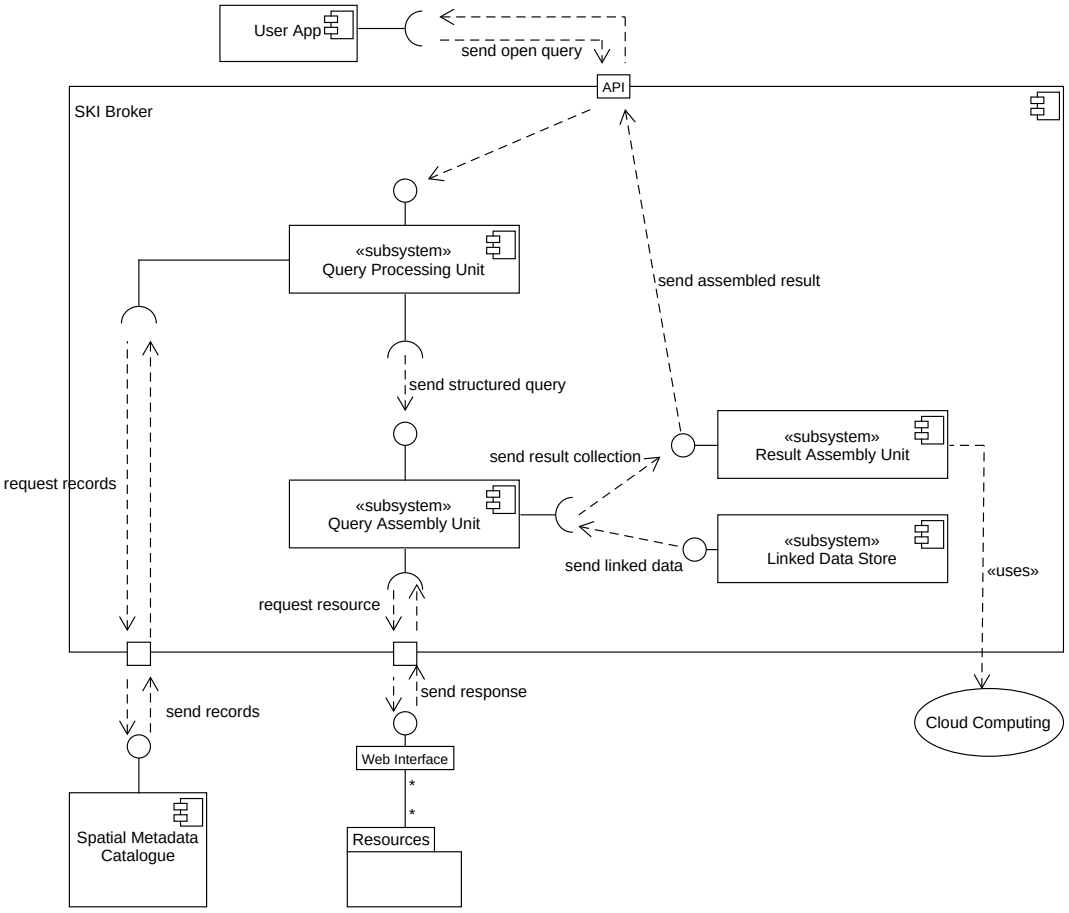

Figure 4: SKI Broker Components - Generalised

As illustrated in figure 4, open queries from the outside world need to be translated into structured query for the broker to understand. This structured query then needs to be assembled and individualised to the disparate but rele-

vant resources on the Web (which include metadata catalogues)—this process utilises a Linked Data Store within the broker to store the semantic representtation of the resources. Once the relevant resources are received, the results are then assembled to provide the solution required by the user. The returned harmonised result can then be displayed by the user apps. Any complex process required by the SKI broker can be facilitated by making use of cloud computing.

For its proper usage, an SKI demands that data be linked. However, this solution depends on an ideal scenario where all data are linked together. To implement an SKI in the current Web, where linked data exist alongside non-linked data, there must be some additional components to transform the non-linked data into linked data. In figure 5, these additional components are shown in red.

An *Ontology harmonization Unit* maps the LD from existing linked data stores to the ontology used by the SKI. This is required because existing linked data might not necessarily use the same vocabulary as the SKI broker, and hence need to be mapped to the SKI's ontology for their proper reasoning and querying.

To allow the reasoning of non-linked data, an *Ontology Creation Unit* is used. It transforms non-linked data into linked data understood by the SKI, allowing their reasoning. The newly created linked data can then be reasoned and queried over.

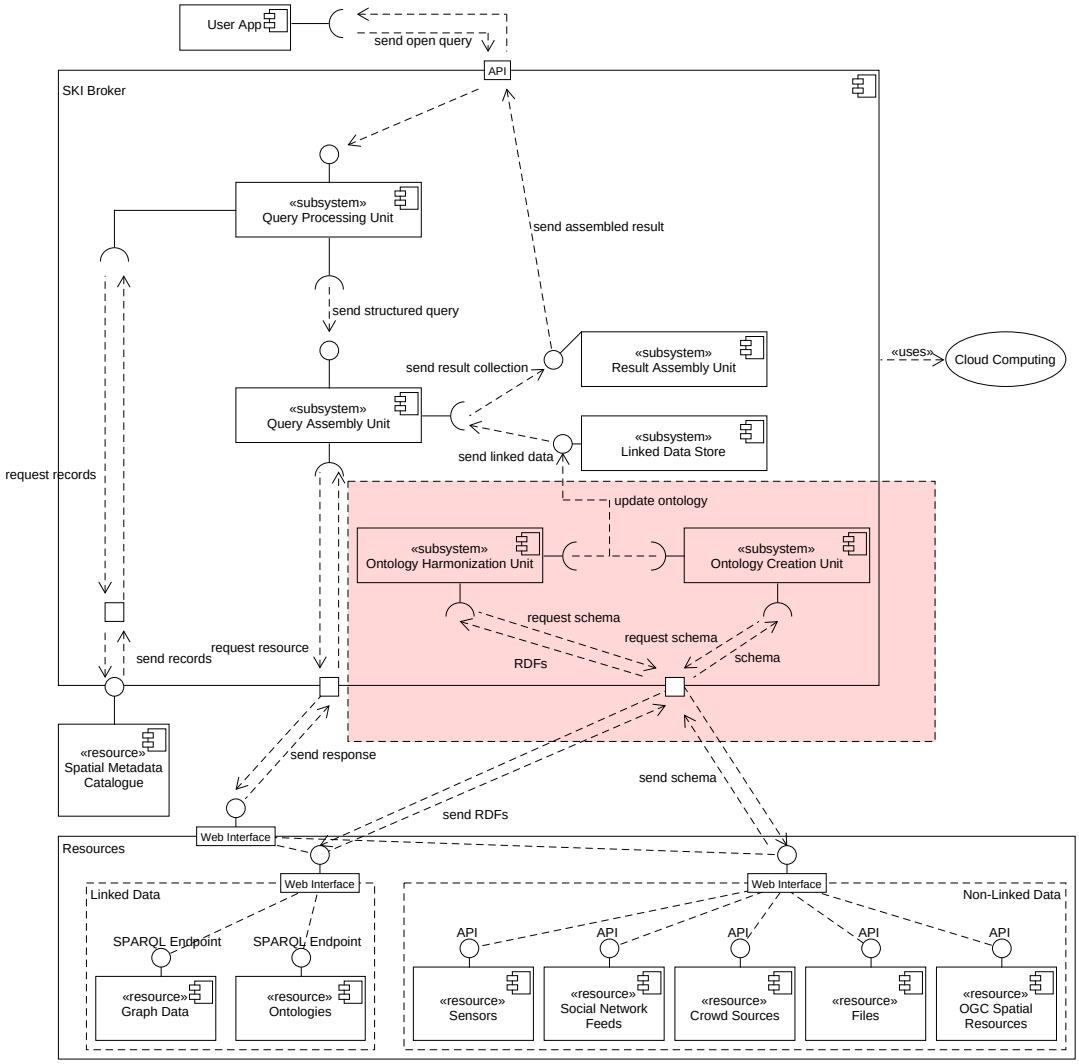

Figure 5: SKI Broker Components - Current

3.4 SKI Implementation

An SKI depends on ontologies where concepts are represented, inferred automatically, and queried over. The main challenge is to change the existing non-linked data into linked data using RDF. While there has been research to (semi-) automatically engineer ontologies (see An and Park (2018), Xiang, Zheng, Lin, and He (2015), Wächter and Schroeder (2010), and Lee, Kao, Kuo, and Wang (2007)), the full automation of this process is unlikely as it is resource intensive, requiring both domain and ontology experts (Simperl & Luczak-Rösch, 2014). A different approach is to utilise an intermediary process that allows LD and non-linked data to interoperate, this is known as Ontology-Based Data Access (OBDA). In this section, both approaches are discussed.

3.4.1 Approaches

Entirely Linked Data Approach

This approach relies on the data providers translating their current datasets into linked data, and enabling a SPARQL endpoint to their ontology. This allows the data providers to keep ownership of their data, and can modify the inferences and rules in their ontology. However, this approach requires all data providers to use the same ontologies and vocabularies to ease the querying of their datasets. While a federated ontology can be used to harmonise different ontologies, this task adds another layer of complexity, and should be avoided. As such, before any spatial dataset is transformed to linked data, the stakeholders of the SKI should agree on the core ontologies to be used at a national level, and domain experts should agree on the inferencing rules needed to facilitate hazard-risk management.

The core ontologies should be robust enough to prevent conflicting knowledge insertion while providing a high-level standard for the representation of the domain. The core ontologies should also be flexible enough to allow data providers to develop a more specialised local ontology based on the core ontologies to best represent their datasets. Finding such a balance requires the participation of all stakeholders involved. For this approach, **federated SPARQL**

queries are possible, and hence a mediator between users and data providers simply need to translate the user's query into federated SPARQL queries to all data providers.

The task of inferencing knowledge is done at the repository level, which distributes the load of such a complex task. However, multiple linked data stores will need to be implemented at each data provider, and there is a paradigm shift that needs to occur (from non-LD to LD). This approach should be the aim of a fully operational SKI, but is time and resource intensive. It requires motivation and commitment to complete this approach, and it should be remembered that this approach is an ongoing effort to be respected by all stakeholders.

Ontology-Based Data Access Approach (OBDA)

This approach makes use of an ontology that maps non-linked datasets to the federated ontology. In this case, there are two ontologies: (1) the federated ontology that represents the domain of interest (e.g. risk hazards), and (2) the intermediary ontology that maps the non-linked datasets to the federated ontology. (1) is used as the base for federated queries, it includes all the rules, inferencing, and axioms required in the domain of interest, and (2) simply acts as a bridge between linked data and non-linked data.

Similar to the previous approach, this one requires core ontologies and vocabularies to be designed. However, the data providers do not need to change their business models and/or datasets. The entirety of integrating (federating, conflating, and fusing) the datasets is done by the SKI broker, which can be expensive. The broker is required to translate the user query into queries understandable by the data providers Web endpoints. Given that they do not have a SPARQL endpoint, different translations might be required; for example, SPARQL to WFS calls, or SPARQL to a specialised API calls. The information required for the translation is stored in the intermediary ontology, and hence high-level inferencing can be achieved (e.g. determining which Web service are more appropriate to answer the user query). However, in this approach, data-level inferencing is difficult. A possible solution for data-level inferencing is to retrieve only the datasets relevant to the user query, add them to the ontology

on-the-fly, inference over it, and answer the query, but this solution is not ideal as the broker's capacity to process simultaneous queries is then limited.

Hybrid

The *entirely linked data* approach requires a lot of time to be fully implemented, and the OBDA approach limits some capabilities of linked data such as inferencing. The hybrid approach aims to combine both approaches, where OBDA is used for data that are not linked, while the *entirely linked data* approach is used for LD stores. During the implementation of an entirely LD SKI, less OBDA processes would be required—easing the processing load on the broker—and, eventually, most of the data will be linked and the *entirely linked data* approach can be used with a minimum amount of OBDA technologies. This alleviates the responsibilities of the broker while still enabling the querying of non-linked datasets. As it is unrealistic to expect all legacy data to be converted to Linked Data (Bikakis, Tsinaraki, Gioldasis, Stavrakantonakis, & Christodoulakis, 2013), this approach is required. Nonetheless, all the approaches require the following components:

1. A core ontology, which is robust enough to act as a standard for the domain, but also allows the development of more specialised versions of it;
2. Agreements from the stakeholders on the approach to use, and how to proceed about it;
3. Ontology and domain experts to guide the stakeholders, and the development of the ontologies; and
4. An SKI broker to mediate between users and data providers.

3.4.2 SKI Technologies

Ontologies

Ontologies for flood-risk management are few, being either incomplete or too generalised. Where ontologies do exist, it is good practice to reuse them, so

promoting global interoperability (W3C, 2013b), and reducing duplication of efforts. While it is acknowledged that most of the ontologies used in an SKI solution for hazard-risk management will need to be engineered, there are nonetheless some existing ontologies that are still relevant for this project which should be considered. A list of these is provided in appendix A.

Linked Data Stores

Linked Data stores are needed to effectively store and query RDF data. There are a wide range of Linked Data Stores available: open-source, commercial, and freemium. Data to be queried in a Linked Data Store should be query-able through a Web service, and be available in standard ontology formats such as Turtle[14], RDF/XML[15], and JSON-LD. For the querying of ontologies, SPARQL is the standard; it allows complex queries to be expressed in terms of triples. A common saying is that SPARQL is for RDF what SQL is for databases. Therefore, the exposed Web Services should enable a SPARQL endpoint to allow for the querying of the linked datasets. While the final choice of a Linked Data Store lies with the implementor of the SKI, a review and recommendation of some them are provided in appendix B of this report.

SPARQL Endpoints

A SPARQL endpoint acts as a bridge between a Linked Data Store and the users. For remote querying, it needs to be exposed *via* the Web. For this reason, Linked Data Stores already implement their own querying interface, and that interface simply needs to be translated into a Web interface with the necessary security features such as sanitisation of the queries. There are specifications for the SPARQL language[16], its protocol specifications[17], and its result specifications[18]. Any SPARQL endpoint implementation can be used as long as they adhere to the standards and meet their design requirements of the SKI. These are left for the software developer to specify.

[14] https://www.w3.org/TR/turtle/
[15] https://www.w3.org/TR/rdf-syntax-grammar/
[16] https://www.w3.org/TR/rdf-sparql-query/
[17] https://www.w3.org/TR/rdf-sparql-protocol/
[18] https://www.w3.org/TR/rdf-sparql-XMLres/

Reasoners

Reasoners are also included in most Linked Data stores, however there are also commercially and open-sourced reasoners available such as Bossam (Jang & Sohn, 2004), RacerPro (Haarslev, Hidde, Möller, & Wessel, 2012), FaCT++ (Tsarkov & Horrocks, 2006), Pellet (Sirin, Parsia, Grau, Kalyanpur, & Katz, 2007), and HermiT (Shearer, Motik, & Horrocks, 2008). These are a matter of implementation based on specific requirements and are left to the software developer to specify.

3.4.3 Global Interoperability

The GSDI Technical Working Group (2009) mentions the importance of an SDI to be interoperable at the global level. The same applies for an SKI—for an SKI to be beneficial at a global level, it should be implemented with the concept of Global SKI in mind. The best way to achieve global interoperability is fundamentally to consider international standards and best practices (Conti et al., 2018). In the case of an SKI, the standards and best practices to be looked at pertain to spatial data and linked data. Section 3.4.4 discusses the existing best practices in regards to spatial data and linked data, while section 3.4.5 discusses the relevant standards.

3.4.4 Best Practices

Publishing Spatial Data on the Web

W3C and OGC have worked together to formulate a set of best practices for publishing Spatial Data on the Web[19]. They are called the Spatial Data on the Web (SDW) Working Group, and have published a set of recommended criteria of best practices (SDW, 2017). A summary produced by the group is found in table 1.

[19] https://www.w3.org/2015/spatial/wiki/Main_Page

Best Practice 1: Use globally unique persistent HTTP URIs for Spatial Things	**Best Practice 8:** State how coordinate values are encoded
Best Practice 2: Make your spatial data indexable by search engines	**Best Practice 9:** Describe relative positioning
Best Practice 3: Link resources together to create the Web of data	**Best Practice 10:** Use appropriate relation types to link Spatial Things
Best Practice 4: Use spatial data encodings that match your target audience	**Best Practice 11:** Provide information on the changing nature of spatial things
Best Practice 5: Provide geometries on the Web in a usable way	**Best Practice 12:** Expose spatial data through 'convenience APIs'
Best Practice 6: Provide geometries at the right level of accuracy, precision, and size	**Best Practice 13:** Include spatial metadata in dataset metadata
Best Practice 7: Choose coordinate reference systems to suit your user's applications	**Best Practice 14:** Describe the positional accuracy of spatial data

Table 1: Best practices for publishing Spatial Data on the Web (SDW, 2017)

Linked Data

Similarly, W3C has published a set of best practices for producing linked data (W3C, 2013a). The seven best practices for producing linked data are listed as:

1. Model the data
2. Name things with URIs
3. Re-use vocabularies whenever possible
4. Publish human and machine readable descriptions
5. Convert data to RDF
6. Specify an appropriate license
7. Host the linked dataset publicly and announce it

5-Star Deployment Scheme for Linked Open Data

Tim Berners-Lee also suggested a 5-star deployment scheme to gauge the quality of Linked Open Data. The criteria is cumulative, where the higher stars presume the previous ones, i.e. to achieve 5-stars, the previous stars from 1 to 4 must also be achieved (W3C, 2013b). This scheme is shown below.

1-Star: Publish data on the Web in any format with an explicit Open License;

2-Stars: Publish structured data on the Web in a machine-readable format;

3-Stars: Publish structured data on the Web in a document, non-proprietary data format;

4-Stars: Publish structured data on the Web as RDF; and

5-Stars: Have the identifiers in the RDF file link to useful data sources.

The first star can be summarised as being openly useable by adding an open license to the data, and being linkable as it must be published on the Web. The

second star refers to the data being machine-readable, meaning using a structured format that allows data to be parsed and processed. The third star ensures that the structured format used is non-proprietary, and hence can be read by any machine free of charge. The fourth star ensures that the linked data format used is internationally standardised, by using the Resource Description Format recommended by W3C, and the fifth star pushes linked data to use other linked data, to promote a Web of linked data.

3.4.5 Standards

Standards are required to ensure interoperability at the syntactic and structural level. That is, by using agreed upon standards, the data make use of the same schema, and the same file format. In a non-linked data environment, Steiniger and Hunter (2012) classify relevant standards into (1) data delivery standards, (2) data format standards, (3) data search standards, and (4) others. Relevant standards for each of these categories in relation to spatial data are mentioned:

1. Data delivery:
   OGC standards for Geospatial Web Services (WMS, WFS, WMTS, WFS-T, WCS, WCPS);

2. Data format:
   GML, KML, WKT, GeoJSON;

3. Data Search:
   CSW, WFS-G, ISO 19115, ISO 19119, ISO 11179; and

4. Others:
   WPS, CTS, WTS, SLD, SE, WMS.

Additionally, there are standards related to the Semantic Web, which are conveyed below.

1. Search and discovery:
   SPARQL;

2. Linked Data format:
   RDF – OWL, XML/RDF, TTL, N3, JSON-LD;

3. Naming things:
   URIs; and

4. Others:
   upper ontologies.

In an entirely linked data environment, only the semantic Web standards would need to be utilised. However, the adoption of the Semantic Web is a slow process, and is always evolving. As such, standards for both non-linked data and linked data need to be considered.

4 Prototype API: A Web-Application for Rainfall Data Analysis

A prototype API was developed as part of this project to demonstrate to flood risk professionals a potential tool which could be of SKI, highlighting some of the key capabilities of such a system, including data linking and near real-time data analytics. The API was developed by Jiakai Li in February 2018 as part of the course *DATA601 Applied Data Science Project*, completed as part of a Masters in Data Science at the University of Canterbury; this section summarises the project report of Li (2018) and presents the outputs from the API.

4.1 Introduction

Rainfall data have become easier to obtain in recent years and are often now readily available online. Using these data, a risk analysis system can be developed for flood or river managers to improve their accessibility to live data and provide near real-time risk information derived from them, enabling quick decisions to be made. The purpose of this project was to establish a prototype of such a system, which can be further developed into a powerful application to acquire and integrate multiple online data sources for comprehensive real-time rainfall data analysis and visualisation. For easy access and duplication, the project was constructed from scratch in a public server (Google cloud) environment, based on open-source statistical computing software of R and its powerful packages.

The web application prototype was developed to conduct the following data analysis tasks:

1. *Data acquisition*: On-demand acquisition of the latest rainfall data from individual gauge observation sites across Canterbury, New Zealand.

2. *Extreme value analysis and mapping data visualisation*: For each site, using the up-to-date rainfall timeseries acquired, a Gumbel statistical probability distribution is fitted to the historical data; the statistical probability of the

most recent rainfall observation is then calculated and visualised in terms of magnitude and likelihood.

3. *Intensity-duration-frequency (IDF) curve analysis*: For individual sites, the web application utilises all the data from that site to calculate and visualise up-to-date IDF curves.

4. *Bayesian inference*: For individual sites, the web application utilises all the historical data and the Markov Chain Monte Carlo (MCMC) method to sample posterior rainfall observations and index parameters, and fit a Bayesian inference model to approximate posterior probability distributions. This enabled uncertainty in calculated rainfall likelihoods to be accounted for, given the variable lengths of the rainfall timeseries available.

4.2 Software Stack

The general software stack for the developed API is shown in Figure 6. The interface between the user and this system is through a website published on Google cloud compute instance. The interaction between the front-end website and the back-end R computing is based on OpenCPU, which could provide a reliable and interoperable HTTP API for data analysis based on R. Moreover, the Bayesian statistical modelling is completed by using PyMC3, a Python package that focuses on the advanced Markov Chain Monte Carlo and variational fitting algorithms. To implement the software stack, several coding developments were needed:

1. Front-end developments were completed in HTML, JS, and CSS, which provided the basic interactions between the web users and the API system.

2. Most parts of the statistical analysis code were written in R, including data import and pre-processing, mapping data visualisation and IDF curve analysis.

3. The Bayesian statistical model was written in Python using packages including PyMC3, NumPy, and TensorFlow (which is a dependency of PyMC3).

The full API was implemented using OpenCPU, meaning that coding for both the website and computing could be integrated into one single R package and modified separately without interfering with each other, thereby providing increased flexibility.

The front-end web interface was developed using:

- HTML and CSS for the basic structure, layout and style, and
- Javascript for the basic functionality and interaction with R via openCPU. Code was developed to implement user interactions, such that user can upload multiple excel files, click on the "update map" button, and also choose site name from the drop bar. From this code, functions in openCPU.js library are called to enable R functions to be run from the back end, and to obtain results. The JQuery.js library was also utilised for efficiency.

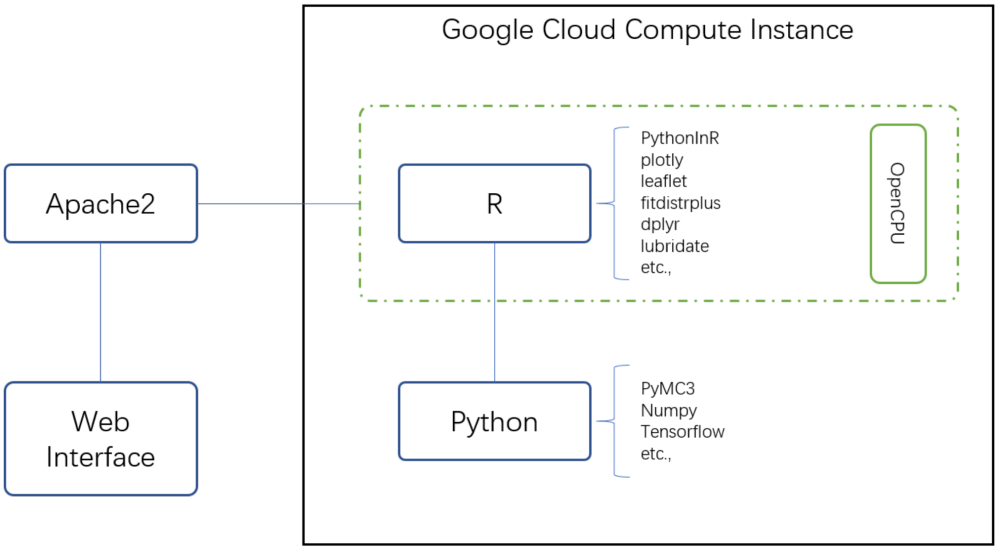

Figure 6: Prototype API software stack

4.3 Approach

4.3.1 Data sources and pre-processing

Rainfall gauge metadata were extracted from Canterbury Maps[20], which provided the site number, name and location (latitude and longitude) of each site. Observation data of raw rainfall values for each site in Canterbury were obtained from Environment Canterbury[21], and included three variables:

1. Observation site number: a unique six-digit code for each site which allows its identification via the metadata, and connects the site to its spatial location enabling mapping.
2. Time of each observation, rounded to the nearest second.
3. Rainfall intensity in millimetres per hour (mm/h) to 1 decimal place. These data were available in comma-separated values (CSV) text format.

The rainfall data files in CSV format were downloaded and imported into a single R data frame object, into which site information (ID code, location and name) for each record was appended. In addition, the text-format time variable was converted to an R time-stamp object, enabling rapid query based on year, month etc. After pre-prosessing, each data entry in the R data frame consisted of six variables: (1) observation site number, (2) time of the records, (3) rainfall intensity records, (4) longitude of the observation site, (5) latitude of the observation site, and (6) the site name, for example:

```
217810 2005-11-01 0 171.8638 -42.75268 MOUNT BYRNE
```

4.3.2 Extreme value analysis and rainfall observations mapping

Rainfall extreme values were analysed based on the Fisher-Tippett-Gnedenko theorem, or extreme value theorem, which deals with the stochastic behaviour

[20] https://canterburymaps.govt.nz/
[21] https://www.ecan.govt.nz/

of extreme values in an event. A probability analysis for the recorded rainfall observation was conducted using a Gumbel distribution, suggested by Nadarajah and Choi (2007) to provide a reasonable model for annual maxima rainfall observations. For each site, the annual maxima rainfall series was extracted and ordered, then the Gumbel distribution fitted using a Maximum Likelihood approach in order to estimate the distribution location and scale parameters with daily updated rainfall observations, the web application was designed to update the estimated parameters every time an update occurs in the input rainfall data files.

The extreme value analysis allowed for the mapping of current daily observations, visualised according to their magnitude and likelihood of occurrence relative to the historical record. Thus, the web application could potentially form part of a real-time risk analysis system, since during extreme rainfall events users can obtain rapid estimates of rainfall probabilities. To produce the map, the latest observation from the rainfall records for each site is extracted, then the cumulative probability for that observation value calculated. Each site was represented on the map using point indicating the site location. The size of the point was determined by the relative magnitude of the rainfall value (against all observations); the colour of the point was determined using the calculated probability, with blues used for frequent (i.e. high probability) observations and reds for infrequent (i.e. low probability, or more extreme) observations.

4.3.3 IDF curve calculation

Intensity-Duration-Frequency (IDF) curves represent a statistical summary of the likelihood of rainfall occurrence at different temporal scales, standardised using the average hourly or daily intensity. IDF curves describe the relationship between rainfall intensity, rainfall duration, and return period (i.e. annual exceedance probability). They allow for the estimation of the return period of an observed rainfall event or conversely of the rainfall amount corresponding to a given return period for various aggregation times (Elsebaie, 2012). Here, the Gumbel distribution was fitted to annual maxima series derived from hourly data for each duration from 1 to 24 hours, using procedures described by Chow

(1951).

4.4 Web application

The web application was developed as a Google cloud compute engine instance[22]. A screenshot of a development version is shown in Figure 7. End users can upload data as multiple CSV files with variables "site no", "DateTime", and "RainfallTotal" (in mm/hr), The raw observation data can be downloaded from Environment Canterbury[23].

In the plotted mapping data (Figure 7, upper section), a bubble on the map stands for an observation site, with a radius proportional to the latest rainfall intensity observation in this site. The colour of the bubble indicates the flooding risk of this observation site, specifically the probability of a higher rainfall intensity than the current one. The colour is divided into three classes according to the probability: (1) high risk (red) means a probability of observing a higher rainfall intensity is lower than 10%; (2) medium risk (orange), means a probability between 10% and 20%; and (3) low risk (blue), means a probability higher than 20%. By combining these two pieces of information, hydrology managers will be able to analyse the rainfall data more easily.

After receiving files uploaded by the web user, the drop bar on the left side of the web page will be updated by unique values from the site names in the uploaded files. The raw data will then be further processed to generate all data needed to plot the IDF curves for the selected site. The web application generates IDF curve return periods of 2, 5, 10, 25, 50, 100, and 1000 years with durations ranging from 1 to 24 hours (Figure 7, middle), and displays the raw rainfall observations as a timeseries (Figure 7, lower). Furthermore, by using the Plotly package, this web application can provide a more interactive figure to help the users interact with the plot. For example, the IDF curves can be zoomed into to view their details more clearly, and an IDF curve for a given return period can be

[22] An updated version of this will be made available via `https://geospatial.ac.nz/rainapp/` in early 2019

[23] `http://data.ecan.govt.nz/`

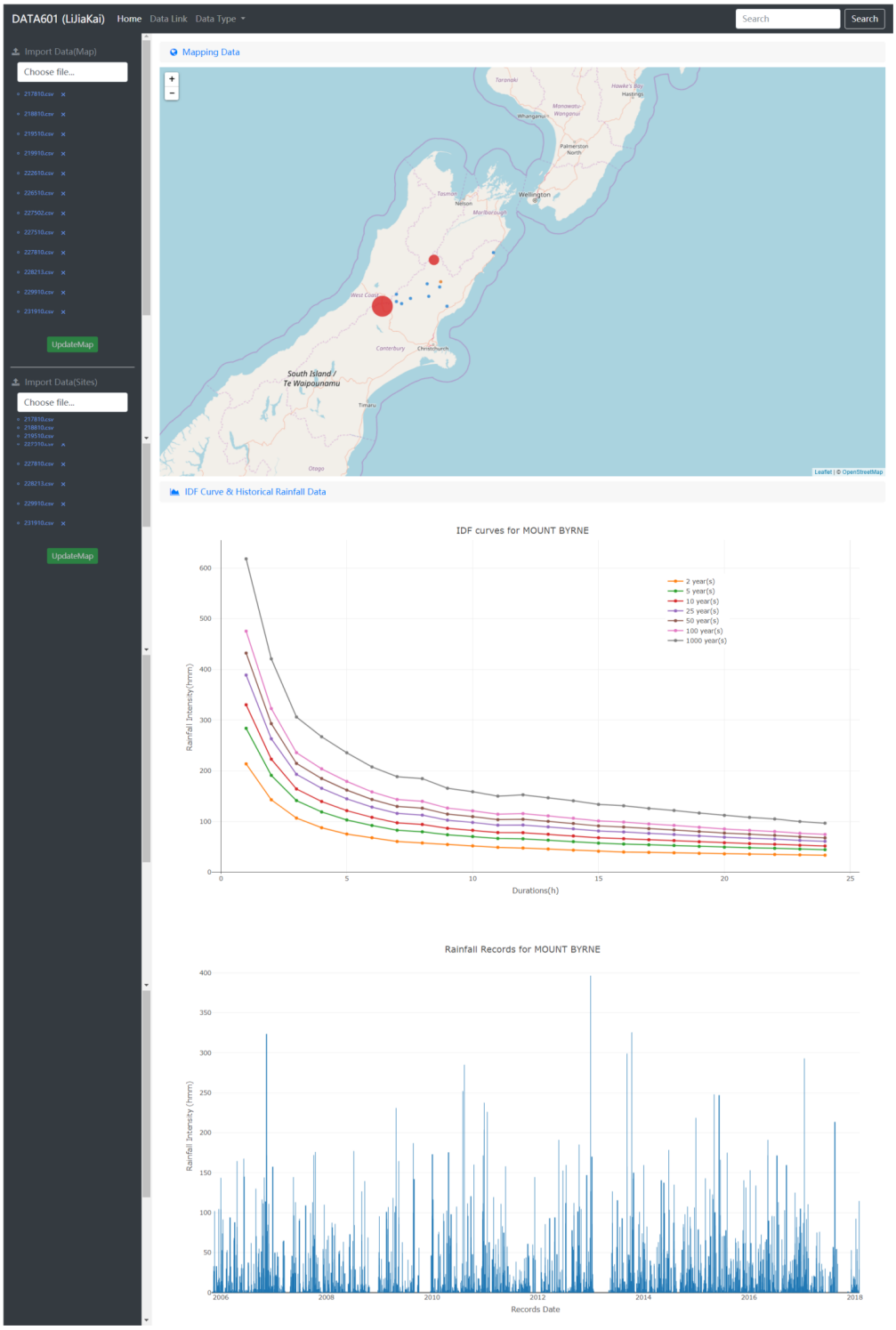

Figure 7: Prototype API website front end

highlighted for selection.

The developed system was illustrated to flood risk practitioners as part of the workshops; results from this assessment are presented in Section 5.3.

4.5 Further developments

The back-end of the website was largely based on OpenCPU, allowing statistical analysis with the R language. However, the inclusion of Python code increased the complexity of the system. For further developments, it would be preferable to avoid using both of these languages. Geospatial libraries are available for both R and Python (e.g. GDAL), but Python is more closely integrated with GIS packages. The statistical analysis functions could be re-written in Python, with server applications provided by Flask instead of OpenCPU to deploy the application: Flask is a comparable package to OpenCPU in Python.

Additional developments of the user interface are also needed:

1. Automatic update of data
   In mapping the data, the data files must be uploaded by the user. The model fits the data each time an update is made. The system can be made to acquire the rainfall data automatically from Environment Canterbury every hour and update the underlying models. This would increase the efficiency of the system for end users.

2. Improvements to IDF curve analysis
   For the IDF curve analysis, this project can successfully produce the curve with a duration ranging from 1 to 24 hours for return periods up to 1000 years. However, to ensure precision, the duration in minutes should be included as well. Moreover, the IDF curves can take various mathematical expressions, but the system was developed with only one expression.

3. Application to other regions
   The system was developed as a prototype to show the feasibility and most functions were designed to process data from Environment Canterbury.

Therefore, the code modification must be undertaken before the system can be scaled to the rest of the regions in New Zealand.

4. Use of additional data sources
   The project only utilised rain gauge data from Environment Canterbury. Additional data sources are available and could be included to provide a more comprehensive system. For example, the web application can be further developed to integrate rainfall observations from data sources such as Global Precipitation Measurement (GPM) from NASA; forecast rainfall could be included through inclusion of weather models such as the Global Forecast System (GFS).

5 Assessment of user-needs: workshops

To understand the needs of flood-risk practitioners, two workshops in Wellington, New Zealand, and Christchurch, New Zealand were conducted in June 2018. The aims of the workshops were to identify the current methods used in flood-risk management, data gaps and limitations, the needs of the stakeholders, and whether real-time analysis and integration of data is a desired and useful step towards better and easier flood-risk prediction and management.

The workshops were divided into three parts. The first part was focused on dataset identification: which crucial datasets are available? what works well in terms of data supply? what are the current related issues? and what improvements can be made to existing practice? The second part was focused on identifying improvements to current tools and desired "next generation" tools. The third part of the workshop was focused on explaining the concept of an SKI, and to gather the users' thoughts on a potential SKI implementation.

Over 60 individuals attended, with representatives from:

- Commercial flood modelling consultancies;
- Crown Research Institutes;
- Regional and Territorial local government;
- Water utility operators;
- Central government departments;
- Regional Civil Defence and Emergency; Management (CDEM) groups; and
- Resilience consultancies.

To help frame the workshops, preparatory interviews were undertaken with a selection of flood modelling practitioners and users to establish typical data requirements, insights into current practices, and barriers/areas for potential improvement in data supply, flood modelling and risk assessment practice.

5.1 Workshop Part 1—Datasets Identification

Part 1 of the workshop focused on identifying the current practices of the workshop participants. This part of the workshop introduced the practitioners to the source-pathway-receptor model of a flood hazard risk, and questions related to the data requirements in these three categories were asked.

The source category relates to data representing or relating to the cause of flooding (e.g. water level), the pathway category is data representing or relating to the conduit of the flood water (e.g. stop banks), and the receptor category relates to data representing or relating to recipients that would be adversely affected by flooding (e.g. people).

The questions asked were:

(1) What are the critical datasets you need for modelling flood risk?

(2) What currently works well or presents problems in obtaining and using these data?

(3) What improvements could be made to this current state, generally, or for specific datasets?

5.1.1 Data Requirements Overview

The responses to the three questions can be summarised into six broad categories:

1. Accessibility;
2. Coverage (spatial and temporal);
3. Governance;
4. Policy;
5. Standardisation; and
6. Uncertainty and data quality.

Accessibility

This category is related to the ease of access of the datasets; whether the datasets can be easily found, downloaded, and used. The responses are summarised below.

- Data download services are available and used to freely access local and central government open data (e.g. `http://opendata.canterburymaps.govt.nz/`, `https://data.linz.govt.nz/`, and `https://data.mfe.govt.nz/`) and provide access to some frequently used datasets.
- The download of large and fragmented datasets is problematic and time consuming.
- Available data is not always easy to discover—making clearer what is and is not available, and what barriers to access and use exists, would be beneficial.
- Greater access to real-time data would be an improvement.

Table 2 shows the answers related to accessibility that were common in all three datasets: source, pathway, and receptor.

Identified Problems	What Currently Works Well	Desired Improvements
Hard to download large datasets	Easy to access download services	Access to data in real-time
Data are saved in individual files, making it slow to download large datasets		Stock-take of existing available datasets and barriers to making them available
Difficult to find and access some datasets (need to know who to ask in the councils)		

Table 2: Accessibility Overview Findings

Coverage

Data coverage relates to the spatial and temporal extent of data available and

whether it is usable, consistent, and fit for use. An overview of the responses related to coverage is given below.

- There is inconsistency, spatially and temporally, in the availability of input data which drives inconsistency in methods and quality of model outputs.
- Datasets with good resolution are becoming increasingly available e.g. LiDAR elevation data or LiDAR derived elevation models.
- Improvements sought are for: greater consistency, coverage and resolution in datasets; the ability to feedback new data into existing datasets to create updates; and centralised repositories for the collation of local datasets.

The overview findings in relation to coverage across all data categories are provided in table 3.

Identified Problems	**What Currently Works Well**	**Desired Improvements**
Insufficient frequency of measurements in datasets both temporally (update cycle and time sries frequency), and physically (spatial resolution)	Resolution of some data	Reduction of current 2-3 hrs time lag in processing real-time satellite imagery
Reliance on averages/Interpolated data due to lack of measure data		Feedback improved DEM's to a central collection.
		Make it easier to share and access model outputs
		Gathering site data and feeding it back for consumption
		Centralised datasets

Table 3: Coverage Overview Findings

Governance

Governance of the data influences its availability and quality throughout the whole data supply chain, including its consistency, and usability.

The responses related to governance are summarised below, with table 4 providing an overview of the responses across all data categories:

- Lack of coordination and agreement on a range of factors (such as standards, method, data discovery and access) is resulting in issues such as duplication, gaps, inefficient pre-processing / reprocessing of data and poor accessibility of data.
- Improvements through better coordination, standardisation, data federation, data management practice and centralised processing have all been identified.

Identified Problems	What Currently Works Well	Desired Improvements
Lack of authoritative source of data; multiple providers/sources of the same or similar data; unidentified data provider		One organisation to collect all data or a high-level coordination of providing consistent data
De-centralised the responsibility for data; it creates inconsistencies at regional and national levels, with gaps and duplication between agencies and their data		Coordinate federated agencies and host data from one place of access
		Super-computers to hold and process data and models
		Establish common responbilities and approaches for dataset collection and management

Table 4: Governance Overview Findings

Policy

Policies ensure that data are available in a consistent and standardised manner. Below is a summary of the findings related to policies.

- There is recognition that both open data policy drivers and commercial imperatives are both at play in the data supply chain, and have an effect on the potential re-use of existing data and model outputs.
- Open data initiatives have been seen to benefit data accessibility and use, and reduce cost.

Table 5 shows the responses obtained that relate to policies across all three categories of source, pathway, and receptor datasets.

Identified Problems	What Currently Works Well	Desired Improvements
CRIs conflicting drivers of providing freely and openly available data and their need to operate on a commercially competitive way and protect their IP	Open and freely available data	
Property data is hard and costly to access—IP is commercially licensed		
Possible ethical concerns about making property values freely available		

Table 5: Policy Overview Findings

Standardisation

Standardisation ensures that data are structured and made available in a consistent manner across different sources. The findings found in relation to standardisation are summarised below.

- Lack of standardisation in several key aspects of flood modelling and data collection is creating adverse impacts in a number of ways:
  - Inconsistent methodologies make comparison of outputs difficult.

  - Inconsistent semantics, data collection methods and processes and datum all contribute to uncertainty in and between model methods and outputs.
  - Lack of consistent data documentation, formats and management all impact on the efficiency and reliability of working with the data and model outputs.

- There are some good examples where standardisation is driving improvement in practice and our ability to model e.g. work done through the LAWA initiative (`https://www.lawa.org.nz/`), national LiDAR base specification (`https://www.linz.govt.nz/data/linz-data/elevation-data`), national horizontal and vertical datum, use of international standards such as WaterML (`http://www.opengeospatial.org/standards/waterml`).

Table 6 shows the responses obtained across all dataset categories in relation to standardisation.

Identified Problems	**What Currently Works Well**	**Desired Improvements**
Lack of common semantics to support users understanding and interpretation of the data	Guidelines for LiDAR, standards for data (e.g. LAWA)	Standardisation of datasets (content, formats, semantics)
Insufficient standardisation in data exchange formats	Interoperability of some datasets	Standardisation of data collection and their coverage
Regional variability of data coverage and quality	National map projection	Need to set common ARI—plus these assessments often vary with hazard type (e.g. volcano, earthquake, flood)
Variation in vertical datums used and the ability to recognise this and interpolate/convert between them	Technology	

Table 6: Standardisation Overview Findings

Uncertainty and Data Quality

Uncertainty and data quality relates to the fit-for-use, and fit-for-purpose of the datasets, and how these can be determined. A summary from the responses are:

- There is a lack of information on data provenance that allows users to assess fitness-for-purpose limits the ability to make informed choices about data use and quantification of uncertainty.
- There is a lack of up-to-date data, or the ability to assess its currency which creates risks of inappropriate use.
- Improved and consistent metadata would help to address some of these problems.
- There is a recognised need to better understand and communicate uncertainty but the methods and tools for doing this do not exist or are not readily or routinely available.

Table 7 shows the responses obtained in relation to data quality and uncertainty across all three data categories of source, path, and receptor.

5.2 Data Requirements Findings

This section reports on the responses obtained in relation to particular dataset categories. Section 5.2.1 reports on the source category, section 5.2.2 reports on the pathway category, and section 5.2.3 reports on the receptor category.

5.2.1 Source

Key source datasets that are necessary to support a range of modelling activities identified by the workshop participants are shown in table 8.

The issues, positive points and potential improvements to the current state of these data are summarised below.

Identified Problems	What Currently Works Well	Desired Improvements
Poor/non-existent metadata describing datasets		Handling of uncertainty
Uncertainty over relevance or fitness-for-purpose of available data		Need to communicate/understand error with different data
Poor warrantability of data; lack of audits to confirm data quality		
Absence of effective understanding and communication of errors and uncertainty in data and model outputs		
Keeping datasets up-to-date (e.g. pipe data)		

Table 7: Uncertainty and Data Quality Overview Findings

Precipitation	Fluvial	Coastal	Tidal Estuarine
Long term rainfall record	River network	Water level–measured	Joint probability water levels
Rainfall–measured	Flows–measured	Water level–modelled	**Groundwater**
Rainfall–modelled	Flows–modelled	Water level–long term record	Groundwater level–measured
Rainfall–probability	Water level–measured	Water level–uncertainty estimates	**Dam Breach**
Rainfall–real-time	Water level–long term record	Wave height–measured	Dams and reservoirs
Rainfall–forecast	Water level–uncertainty estimates	Wave height–modelled	**Tsunami**
Rainfall–forecast uncertainty	Water level–modelled	Wave height–forecast	Models
Climate change scenarios	Water bodies		Recorded

Table 8: Source Critical Datasets

Issues

- Combining data from multiple providers / source impacts reliability and adds time.
- Spatial distribution of measuring sites (rainfall, water levels, waves) is too sparse—impacts on forecast and model reliability.
- The lag between real-time events and data availability is too great for near real-time modelling.
- The need to interpolate between modelled outputs introduces greater uncertainty.
- Gaps in key datasets–spatially, temporally (real-time and long-term record)–even when model inputs are patched together from different sources (authoritative and crowd sourced).
- Particular shortage of water level measurement sites in estuaries.
- Poor rain and snow data at high elevations plus remote upstream areas.
- Reduction in the number of groundwater monitoring sites resulting in a shift from use of data real-time to averages.

Positives

- Data is available from council website / services.
- NIWA river network and High Intensity Rainfall Design Service (HIRDS) availability.

Improvements

- Fine resolution for small catchment assessment.
- Gauge-corrected rain radar to provide real-time data feeds.
- Improved rainfall radar to provide both raw data and rainfall estimates with confidence estimates.
- Disaggregation of long-term sea level resolved in to different components (astronomic surge, tsunami etc.)

5.2.2 Pathway

Table 9 shows the critical datasets identified that are necessary for a range of modelling activities. Issues, positive points and potential improvements of the identified datasets are provided hereafter.

Issues

- Land use classes used to infer surface roughness can be too coarse.
- Currency: land surface has changed meaning DEMs are out of date; storm water network changes are not recorded.
- Inconsistency in use of available vertical datums.
- Inconsistent resolution and accuracy in DEMs.
- Slow uptake of the use NZ Vertical Datum 2016–requires more conversion grids.
- Topographic data is poor where there is no LiDAR coverage.
- Access to existing topographic surveys.

Ground and Land Cover Characteristics	Channel and Floodplain Characteristics	Infrastructure	Tsunami
Soil moisture	Ground surface representation: Digital Elevation Model (DEM), Triangulated Irregular Network (TIN)	Stopbanks: location, construction, condition, crest level (design v actual), consented and unconsented	Joint onshore–offshore DEMs at appropriate resolution
Surface roughness	Channel bathymetry and capacity	Road and rail bridges with levels	Offshore bathymetry grids
Surface permeability	River/floodplain cross-sections	Storm water network components and capacity	River bed DEMs
Infiltration capacity/losses	Average Recurrence Interval (ARI)–for protection and out-of-bank flow	Surface features/impediments: embankments, buildings, walls etc.	River water levels
Evapotranspiration	Raw and processed data for topographic and hydrographic survey	Culverts	Onshore roughness
Soil type	Water bodies		
Land cover	Flood storage areas		
Vegetation	Historic flood outlines		
	Catchment boundaries		
	Overland flow path		
	Time to inundation		
	Shoreline including river mouths		

Table 9: Pathway Critical Datasets

- Information on stopbanks is inconsistent, incomplete and often not current. Unconsented stopbanks are often unrecorded.
- Poorly defined and out-of-date and data availability of shoreline/coastline reference positions, including estuaries/river mouths.
- Topography and hydrography is dynamic (e.g. earthquakes, slips, construction, floods and erosion/deposition).
- Drains and streams not accurately or consistently defined in drainage network datasets.
- Numerous incompatible and variable datasets of land cover and soil datasets across NZ.

Positives

- Ground model data posted to OpenTopography.Org.
- High quality openly available LiDAR where it exists.

Improvements

- Capture and publish flood action plans activation levels.
- Standardised procedures to assess hydrology/runoff.
- Standardised freeboard calculation guidelines.
- Use of national standard for collecting, processing LiDAR accuracy to precision .

5.2.3 Receptor

Critical datasets identified in relation to the receptor category are provided in table 10.

People	**Property**	**Critical Infrastructure**
Statistics NZ census datasets	Building outlines	Emergency routes, evacuation corridors
Statistics NZ Integrated Data Infrastructure (IDI)	Building use/type and value/cost	NZ Lifelines national infrastructure vulnerability report
Social vulnerability index	Surveyed floor levels/threshold levels	Asset value
Dwellings	EQC land damage claims data (flooding)	Asset condition/vulnerability
Critical customers	District planning data/maps	**Environmental Assets**
Rest homes	Property value	Ecological datasets
	Damage and clean-up costs	**Culture and Heritage**
	Consents	Council/culture heritage registers
	Current land use and value	
	Future planned land use/urban growth/zoning maps	
	Debris surveys	

Table 10: Receptor Critical Datasets

The issues and potential improvements to the identified pathway datasets are shown in table 11. No positives were identified by the workshop participants.

Issues	Improvements
Assumed population figures based on average numbers per dwelling.	Develop tools/scripts to make estimates available online.
Assessing history of floodplain development.	Days under water–pasture damage i.e. how quickly the paddock drains, advice to district councils.
Availability, accessibility and re-usability for any required receptor dataset.	Floor height, feed national model.
Infrastructure–no data on some assets, no common model for data and analysis, collating different asset and utility types of data, attributes/ metadata varies, indirect damages, intangible losses, clean-up costs.	Census data - spatially available: average household size, demographics.
Costs/ damage–inaccessibility of insurance of datasets, insufficient granularity/resolution in data, poor accessibility of asset information.	Key assets and lifelines database.
What constitutes environmental damage? Little information available (except from DoC).	Updated topographic survey information could be used to update asset data.
Floor levels–minimum information on existing floor so these have to be surveyed; lack of accuracy standards for floor level survey.	Building consents data made available.

Table 11: Receptor Issues and Improvements

5.3 Workshop Part 2 - Tools Identification

The second part of the workshop was aimed at gaining an insight into which tools would be useful and practical for flood risk practitioners, for either planning or emergency management. This section started by briefly defining a SKI as *"a network of data, analytics and policies allowing data integration and analysis in real time"* (further detail on this was provided in Part 3 of the workshop, see Section 5.4). Two examples of analytical applications which can be developed as part of a SKI for flood risk were then illustrated:

- Statistical analysis time-series data in real-time: a working demonstration as detailed in Section 4, and
- Advanced stochastic flood modelling for scenario assessment: a conceptual outline, as summarised in Figure 8.

While viewing the demonstrations, participants were asked to consider which tools they would find to be most useful for flood risk management. Following the demonstration, the specific questions addressed were:

(1) What tools would be most useful? What is a priority? What's on your wish-list?

(2) Which methods can be further developed?

(3) What linkages can we make to other tools or models? (e.g. RiskScape? Climate models?)

(4) What would be the use of depth/ probability outputs c.f. single event simulations?

Participants addressed these questions simultaneously in their discussions. Sections 5.3.1 and 5.3.2 deal with the tools requirements and issues identified, respectively; 5.3.3 then comments on the potential for the development of these tools.

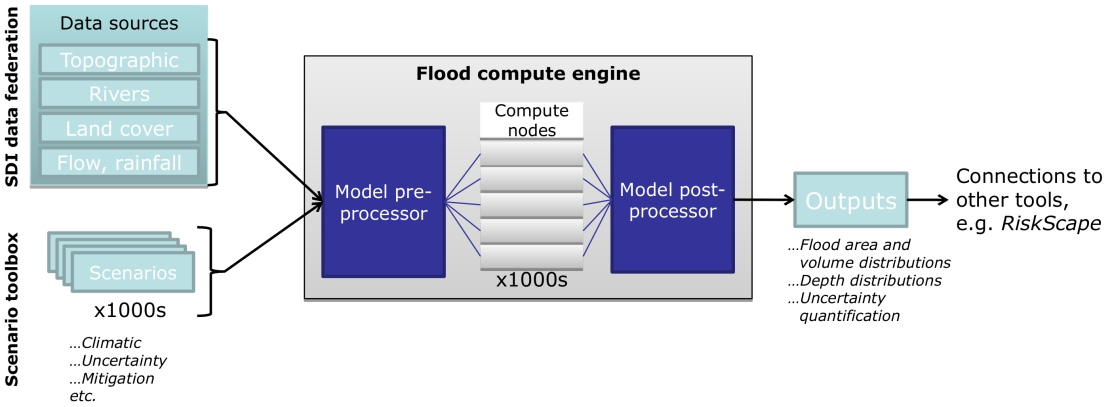

Figure 8: Conceptual outline of an advanced stochastic flood modelling system. Users were presented with the concept of using a SKI to assist with model development and facilitating the running of many 1000s of scenarios via cloud computing. Such scenarios include detailed assessment of exceedance probability, uncertainty analysis of model simulations, or "what-if" scenarios such as potential stopbank (levee) breach or altered flood risk due to climate change.

5.3.1 Tools

Priorities

The development of a national flood model system was recognised to be a priority for development. This would consist of hydrology tools including rainfall/ runoff and river models and an associated toolbox for data processing and analysis of model outputs. It was noted that the system should:

1. Be consistent across the country and provide a first cut at flood information where no existing model exists;

2. Have flexibility, including an ability to upload and run existing flood models and utilising a flexible mesh;

3. Answer fundamental flood risk questions as a priority, including: where floods, how often, when, how deep, for how long, and when will the flood recede?, taking into account protection requirements such as stopbanks and drains;

4. Provide outputs as depths with associated probabilities and allow data export and validation; and

5. Integrate with other applications and systems, such as a SDI which can be developed to provide national flood data, or BIM for asset data, or rainfall probability models; this will allow the systems developed to consume boundary conditions which utilise and respond to local conditions.

It was also recognised that a complementary approach between fluvial and pluvial flooding is needed, with respect to a combination of experience and expertise needed to building a flood model system, and with regard to modelling scenarios.

Scenarios

It was suggested that the system developed should be used for scenarios which:

1. Facilitate impact evaluation, rather than only focussing on hazard modelling, to allow assessment of flood consequences for given scenarios through integration with "receptor" datasets;

2. Enable the development of high-detail and site specific models for infrastructure design activities such as land-use planning, or for the development of mitigation scenarios;

3. Allow scenario modelling, including scenarios which explore differences in volume above critical thresholds vs peak water level and discharge (which are derived from assessments of annual exceedance probability), or changes in rainfall intensities such as derived from climate models or the High-Intensity Rainfall Design System (HIRDS)[24] developed by NIWA; and

4. Ideally allow near real-time modelling, on-the-fly during an emergency situation, or now-casting based on rain gauges or forecasts, with depths provided in terms of building floor levels.

[24] https://hirds.niwa.co.nz/

The recognised importance of impact evaluation as an output of the modelling system suggests that a dynamic integration with tools such as RiskScape[25] would be beneficial. It was also stated that in order to achieve national consistency, the model system will depend on standardisation, particularly in terms of data collection; it was suggested that a tool could be developed to help prioritise future data collection at a national scale.

5.3.2 Issues

System end users

Questions were raised regarding the identity of the system end users, in particular with regard to whether they might be members of central or local government, or consultants. However, it was thought important that it should be made easier for smaller groups to set up localised models from the national dataset/ tool. These groups may include local or regional councils, or local communities.

Communication and model uncertainty

The communication regarding outputs and modelling was highlighted as being critical for appropriate interpretation, and it was suggested that public authorities may be able to contribute funds towards the development of appropriate visualisation tools, such as through using augmented or virtual reality systems (AR/VR). The clear communication of modelling assumptions, the underlying models and data used, were highlighted as crucial to avoid possible mis-interpretation. In particular, developing an effective strategy for the communication of model uncertainty was recognised as important and it was suggested that model scenarios should include an assessment of how uncertainties propagate and effect model results.

[25] https://www.riskscape.org.nz/

5.3.3 Potential

The identified priorities suggests a strong support within the professional community for the concept of an SKI, with a clear desire to make flood modelling and flood risk assessment more accessible and routine. However, concerns were raised regarding potential issues with regard to local modelling vs. national consistency, particularly since supporting data are not yet available nationwide. Scale issues have been identified as likely to be a significant challenge, particularly with respect to the automation of detailed, large-scale localised models in urban areas, which need, for example, to include detailed information on surface and underground drainage infrastructure. The automation of such models is likely to be difficult, and a multi-scale modelling approach may be required, such as through the use of a nested system of models.

5.4 Workshop Part 3 - Imagining the future of SKI

The purpose of part 3 of the workshop was to identify whether the stakeholders related to a flood-risk hazard management found value in the use of an SKI. The following three questions were asked:

(1) Data access: How do you currently access flood modelling data?

(2) SKI benefits: How can an SKI benefit you? and

(3) SKI concerns: What are your concerns regarding an SKI?

It is to be noted that the concept of an SKI was not familiar to the workshop attendees. As such the questions asked were aimed at providing clarification of current practice and confirmation of assumed benefits of using an SKI rather than exploring the features and working of an SKI in technical detail.

5.4.1 Data Access

The data access question was aimed at gauging the current practices in regards to obtaining flood-related data. From the responses obtained, current access to data can be categorised as council websites, desktop applications, personal contacts, and other custom APIs from various organisations.

From these, it could be observed that no centralised access to related data could be found in any one place, and that it is difficult to find these pieces of information. More surprising is that some datasets are obtained by knowing a particular contact person and phoning or emailing them to get a download link to the datasets.

These answers show that there is a need for a centralised method to search for and retrieve datasets irrespective of their formats and locations.

5.4.2 SKI Benefits

The question posed in this section aimed to build understanding of where the stakeholders saw value in an SKI. The responses obtained can be categorised as:

- **Data consistency:**
  Greater data consistency was seen as positive in an SKI infrastructure because the data can be standardised using the Semantic Web and ontologies.
- **Data accessibility:**
  The accessibility of the data was seen as being easier in an SKI given that data will be linked, and hence no 'islands' of data would exist.
- **Data understanding:**
  Given that the purpose of ontologies is to map the knowledge in a machine-readable manner, the data in an SKI would be more understandable as they would be formally described with ontologies. Further, knowledge can be easier understood, and less technical skills are required as they would be mostly machine processable.
- **Data exploitation:**
  Understanding the data leads to better exploitation of the data. For this category of responses, some use of data in an SKI were: real-time analysis of the data (as soon as they get uploaded), predictions using the data, Artificial Intelligence (AI) related tasks, and the enablement of more expressive queries.
- **Data customisation:**
  Providing a common foundation to express knowledge and concepts in an SKI led to the stakeholders suggesting that the manipulation of the data outputs would be easier. For example, flood maps using disparate satellite photos, and data translations that are customised to the user.
- **Data confidence:**
  The SKI can provide a layer of confidence that is based on collective expert effort to formalise concepts use and knowledge created. This ensurers that

the chosen concept is modelled by experts in that domain. Further, the data used can be traced back if knowledge provenance is included in the SKI. This means that the trust level and confidence of the data can be confirmed to the user.

5.4.3 SKI Concerns

Apart from the challenges that are present in implementing an SKI (e.g. organisational, funding, capability) other concerns raised by the stakeholders relate to its automation. While having machines automatically process the data can be efficient, there are concerns in regards to the trust of the inputs, and processes; the quality of the output depends on the input. Further, there were concerns in regards to the provenance of, not only the data, but also of the processes and modelling stages of the SKI—how can the processes be confirmed and trusted if they are all automated.

In addition, by moving more manual processes to automated ones, the issues of ethics and morals need to be addressed. The stakeholders mentioned that some knowledge can be detrimental to particular parties, but in an SKI, such knowledge would be available to anybody, and could potentially be misused by anybody. Another aspect is the misinterpretation of the data by the users; while experts traditionally can explain what the data means, in an SKI, the user may depend solely on their own interpretation of the ontologies, modelled concepts and outputs, hence data and knowledge representation is a concern to address.

In regards to the implementation of an SKI the issues of funding and modelling were raised. An SKI requires a lot of ongoing collaboration among experts, and funding would be required to maintain a national SKI. Hence, an SKI is resource intensive at the initial stage both in human and monetary resources.

In summary, the main concerns raised were about trust of the automation process, the open data model paradigm, the user interpretation of data, and the resouces required to implement and maintain an SKI.

6 Discussion

There is clear demand, particularly from local government and utility operators, for flood modelling services. Typically these are to support local plans, flood protection scheme design and evaluation, and asset and infrastructure design and management. This demand is met by a mature and well skilled supplier community based within the private, research and local government sectors. Models tends to be created to simulate specific events (e.g. for a particular probability of occurrence / return period) rather than establishing flood / depth probabilities. Modelling to determine quantified risk (as a function of probability and consequence) is not common.

6.1 Key findings

There is a clear need for improvement in the current SDI. In this project's use case—hazard risk management—the lack of data consistency, quality, and their real-time feed were identified by the workshop participants to be in need of improvement, alongside their accessibility and documentation. While the participants have identified some positives to the current datasets, only a few selective datasets were mentioned in this regard. Out of the findings, it can be summarised that the participants desire data that is of high quality, consistent, easily accessible (e.g. centralised access), and of higher coverage. Out of the three data categories (source, pathway, receptor), the category requiring more improvement can be said to be the receptor dataset, which requires more data to be surveyed and made available.

A main tool identified to be a priority was a national flood model system. The participants desired a national flood model system that would be consistent, and provide live flood information. However, such a system should be open so that experts can input their own datasets, and extract their needed information. Such a system would help in facilitating impact evaluation of a flood, enable the development of site specific models for infrastructure design activities, and allow near real-time scenario modelling. It was also recognised that the flood

model system should be flexible enough to allow its integration in existing risk assessment tools such as RiskScape.

In terms of issues with current flood tools and systems, the uncertainty and visual interpretation of flood datasets were raised. It was agreed by the participants that more work was needed that would allow users to correctly interpret flood datasets in ways that recognise and account for uncertainties. Virtual reality and augmented reality were suggested to be a possible solution.

The identified priorities for improved flood-risk management and the issues identified with the current system potentially lend themselves to resolution through the implementation of a Spatial Knowledge Infrastructure (SKI). Better data accessibility, consistency, interpretation, exploitation, customisation, and confidence could theoretically be achieved with an SKI. The participants all found benefits in the concept of an SKI for better flood-risk management. However, there were concerns identified, mainly of trust of the automation process, information misuse, data interpretation and of the funding of such a system.

None of these are easily overcome, therefore small-scale or partial demonstrators are necessary as a first-step to provide tangible, beneficial outputs and confidence that an larger SKI implementation would be worth investing in.

6.2 Pathways forward for SKI implementation

To fully implement an SKI, the participation of stakeholders is crucial. We have taken this first step by gauging the interests of flood-related stakeholders, and creating a group of users that can engage with this project as it develops to provide us with essential feedback.

For the next phase demonstrator use cases for the SKI will need to be identified alongside the requirements that the system needs to fulfil. A grant will be obtained to develop an SKI of small scale for a particular use case.

This should be open source and simple enough for easy adoption, as well as being independent of a single platform/server to host it on (as this incurs ongoing costs). For this aspect, the SKI should be mobile enough for any organisation

to deploy their own‘mini-SKI’ on existing infrastructure, while allowing regional councils and aggregation parties to adopt an SKI that federates those mini-SKIs. Hence, a ‘web of SKIs’ will be produced, enabling any SKI holders to interact with any other linked SKI.

6.2.1 User Engagement

An SDI’s primary aim should be user adoption (Hendriks et al., 2012; Masser, 2017), this aim applies to an SKI also. A spatial infrastructure should not be seen as a stand-alone tool to be used, but rather as a foundation that is assimilated by a user to facilitate a more homogeneous experience with spatial data—hence the term infrastructure. While the functional aims of an SKI are important, Hendriks et al. (2012) state that an infrastructure has better chances of serving its purpose if user adoption is kept in mind. Further, involvement with the stakeholders is a needed step to ensure user needs are addressed (Baker, Coaffee, & Sherriff, 2007; Kmoch, Klug, Ritchie, Schmidt, & White, 2016), as well as to facilitate the dissemination of information (Conti et al., 2018). User engagement is as important in ontology design, where the stakeholders need to agree on requirements, priorities, and about alternatives to representing the domain concepts for the interests of both the individual and the community (Simperl & Luczak-Rösch, 2014).

To engage the users, there must be ongoing workshops and presentations in regards to the implementation of the prototypes. Doing so creates the necessity to use technologies and components that can be adopted by the users (Kmoch et al., 2016). Moreover, to facilitate data sharing (whether linked or not), Wallis, Rolando, and Borgman (2013) stress that existing social and cultural implications of the users need to be considered.

Therefore, we suggest that stakeholders engagement workshops and presentations be held at regular interval during the implementation of a flood SKI. Doing so will result in an urgency for frequent updates and prototyping, which will drive a user-focused implementation of the SKI, resulting in a higher likelihood of adoption.

6.3 Challenges to SKI technical implementation

While an SKI can resolve interoperability issues and allow for real-time processing of data, there are various challenges to be addressed.

Implementing an SKI carries many of the challenges of SDI implementation e.g agreeing governance, policies and consistent technical approaches at a national level with multiple interested parties are very difficult and are often the reasons behind slow or incomplete implementations. With an SKI the additional need to develop a semantic enablement layer supported by linked data will be a significant challenge, and though linked data approaches have been used for over a decade they are not mainstream to most organisations producing and using the types of data identified as key to flood risk modelling.

URI Patterns and Minting

Linked data rely on naming each 'thing' using a persistable and unique URI. While any unique and persistable URI can be used, there needs to be a common agreement in regards to the domain, subdomains, and path patterns of the URI. This is needed for the proper scaling and maintenance of linked data (Yu & Liu, 2015); as more data graphs are named, there are less URIs available, and hence well-designed URI patterns and naming conventions are needed to cater for all abstract levels of 'things' ensuring no conflicts during the naming process. However, the same issue of participation remains, where the responsibility of properly using the conventions specified still rests on the user, or on a moderating body.

In addition, a graph whose URI changes will directly affect existing applications that are dependent on the graph. Resolving this issue requires a process known as URI minting. This process involves using a URI that can be redirected to other URIs. As such, if a URI needs to be changed, the minter URI will reflect that change. For this to happen, a common domain needs to be managed by a centralised body, for example, PURL[26], and the URI patterns described previously need to be respected by the centralised body.

[26] `https://archive.org/services/purl/`

SPARQL

SPARQL endpoints are gateways that allow the querying of linked data through SPARQL. An endpoint, however, requires the exposure of the Web service which leads to various security vulnerabilities (Kumar & Kumar, 2014) that need to be addressed. For example, malicious SPARQL injections are possible (Bamashmoos, Holyer, Tryfonas, & Woznowski, 2017), and given that the time complexity of SPARQL is polynomially related to the input query (Perez, Arenas, & Gutierrez, 2006), a Denial of Service attack can be easily achieved. Therefore, the types of queries that the endpoint allows need to be restricted, or filtered (as in Bamashmoos et al. (2017)), or ontologies can be used to detect these Web attacks (Razzaq et al., 2014). However, the issue remains of deciding the threshold timeout value, or the type of query where less expressivity could lead to more vulnerability, but a more secure endpoint reduces the querying potential of linked data.

Yu and Liu (2015) mention that composing SPARQL queries is difficult for the common user as they need to first fully understand the queried ontology. Instead, Yu and Liu (2015) suggest that linked data should be provided in a RESTful manner. However, this would limit the capabilities of SPARQL, hence a compromise needs to be met to ensure expressivity against ease of use.

SKI Ontologies

While the adoption of semantic Web technologies is growing (Bikakis et al., 2013; Schmachtenberg, Bizer, & Paulheim, 2014), ontologies related to the spatial domain are still lacking, being either incomplete or too general. Even though related ontologies are noted in appendix A, these ontologies are to be considered as potential parts of a greater ontology whose core ontology is for hazard-risk management. This project still requires major engineering of ontologies for both spatial data and flood-related data. In addition, organisations will also need to follow standards and conventions imposed by Linked Data and RDF to properly use the SKI and its ontologies.

Ontology Evaluation

What constitute a good ontology is still the debate of much research (Amith, He, Bian, Lossio-Ventura, & Tao, 2018; Lourdusamy & John, 2018). However, a main success criterion is ontology usability—whether the ontology can be used by specific users to achieve goals effectively, efficiently, and satisfactorily (Ma, Fu, West, & Fox, 2018). Nonetheless, this gap in the semantic Web research makes it difficult to properly choose or evaluate ontologies to be used or engineered for the SKI.

Data Licensing

The semantic Web is based on the promise of open data, where all data are shared, and hence can be openly accessed. There are available license ontologies (W3C, 2017a) that can be used, and Schmachtenberg et al. (2014) found that the *dct:license*, *cc:license*, and *dc/dct:rights* are the most important RDF terms used. Nonetheless, there are still issues to be addressed such as (1) restricting access to Linked Data, (2) charging for the use of commercial Linked Data, and (3) charging for data used in deriving facts (Arnold et al., 2018).

Semantic Web Evolution

The semantic Web, like any technology, has undergone many evolutionary steps throughout the years starting with RDF (W3C, 2014a) and RDFs (W3C, 2014b) to OWL (W3C, 2004) and OWL2 (W3C, 2012), and SHACL (W3C, 2017c) and ShEx (W3C, 2017b) for validation. More changes are likely to occur and as such, during the development of an SKI, it is important to keep up-to-date with these changes. Reed (2017) mentions that INSPIRE underwent a similar development cycle as for software, and this is likely to happen during an SKI implementation.

7 Conclusion

The findings from workshops in Wellington and Christchurch with flood-risk management practitioners indicate that the implementation of a Spatial Knowl-

edge Infrastructure can provide solutions to some of the problems faced regarding data accessibility and analysis. Improvements made though taking an SKI approach include better data accessibility, consistency, interpretation, exploitation, customisation, and confidence.

The second phase of this project will focus on a small scale SKI prototype specific to a particular flood use case. Afterwards, it is hoped that the prototype will promote the implementation of a large scale SKI that can be used for hazard-risk management. It is planned that the SKI will be open sourced after the funding runs out, and that the SKI will be distributed in nature. That is, it will not run on a centralised system but rather, different data providers and users can run their own SKI, hence producing a Web of SKIs.

As such, this report has introduced the concept of an SKI and has laid down the starting block to implement an SKI for hazard-risk management. The standards, tools, technologies, and architecture were outlined and reviewed. Alongside these, the challenges to address were mentioned and the next steps towards a full implementation of an SKI for better hazard-risk management were suggested.

References

Amith, M., He, Z., Bian, J., Lossio-Ventura, J. A., & Tao, C. (2018). Assessing the practice of biomedical ontology evaluation: Gaps and opportunities. *Journal of Biomedical Informatics*, *80*, 1 - 13. Retrieved from http://www.sciencedirect.com/science/article/pii/S1532046418300285 doi: https://doi.org/10.1016/j.jbi.2018.02.010

An, J., & Park, Y. B. (2018). Methodology for automatic ontology generation using database schema information. *Mobile Information Systems*, *2018*.

Arnold, L. M. (2016). Improving spatial data supply chains: Learnings from the manufacturing industry. In *The eighth international conference on advanced geographic information systems, applications & services.* Venice, Italy.

Arnold, L. M., McMeekin, D. A., Ivánová, I., & Armstrong, K. (2018). *Knowledge on-demand: A function of the future spatial knowledge infrastructure.*

Baker, M., Coaffee, J., & Sherriff, G. (2007). Achieving successful participation in the new uk spatial planning system. *Planning Practice & Research*, *22*(1), 79-93. doi: 10.1080/02697450601173371

Bamashmoos, F., Holyer, I., Tryfonas, T., & Woznowski, P. (2017). Towards secure sparql queries in semantic web applications using php. In *2017 ieee 11th international conference on semantic computing (icsc)* (Vol. 00, p. 276-277). Retrieved from doi.ieeecomputersociety.org/10.1109/ICSC.2017.29 doi: 10.1109/ICSC.2017.29

Bates, P., & Roo, A. (2000). A simple raster-based model for flood inundation simulation. , *236*, 54-77.

Berners-Lee, T., Hendler, J., & Lassila, O. (2001). *The Semantic Web.* Scientific American. Retrieved January 15, 2018, from http://web.cs.miami.edu/home/saminda/csc688/tblSW.pdf

Bikakis, N., Tsinaraki, C., Gioldasis, N., Stavrakantonakis, I., & Christodoulakis, S. (2013). The xml and semantic web worlds: Technologies, interoperability and integration: A survey of the state of the art. In I. E. Anagnostopoulos, M. Bieliková, P. Mylonas, & N. Tsapatsoulis (Eds.), *Semantic hyper/multimedia adaptation: Schemes and applications* (pp. 319–360). Berlin, Heidelberg: Springer Berlin Heidelberg. Retrieved from https://

doi.org/10.1007/978-3-642-28977-4_12 doi: 10.1007/978-3-642-28977-4_12

Brown, J. D., & Damery, S. L. (2002). Managing flood risk in the uk: towards an integration of social and technical perspectives. *Transactions of the Institute of British Geographers*, *27*(4), 412-426. doi: 10.1111/1475-5661.00063

Chow, V. T. (1951). A general formula for hydrologic frequency analysis. *Eos, Transactions American Geophysical Union*, *32*(2), 231–237.

Conti, L. A., Filho, H. F., Turra, A., & Amaral, A. C. Z. (2018). Building a local spatial data infrastructure (sdi) to collect, manage and deliver coastal information. *Ocean & Coastal Management*. Retrieved from http://www.sciencedirect.com/science/article/pii/S0964569117306634 doi: https://doi.org/10.1016/j.ocecoaman.2018.01.034

Dottori, F., Martina, M., & Figueiredo, R. (2016). A methodology for flood susceptibility and vulnerability analysis in complex flood scenarios. doi: 10.1111/jfr3.12234

Duckham, M., Arnold, L., Armstrong, K., McMeekin, D. A., & Mottolini, D. (2017). *Towards a spatial knowledge infrastructure* (Tech. Rep.). Australia and New Zealand CRC for Spatial Information. Retrieved May 24, 2018, from https://www.crcsi.com.au/assets/Program-3/CRCSI-Towards-Spatial-Knowledge-Whitepaper-web-May2017.pdf

Elsebaie, I. H. (2012). Developing rainfall intensity–duration–frequency relationship for two regions in saudi arabia. *Journal of King Saud University-Engineering Sciences*, *24*(2), 131–140.

EM-DAT. (2018). *Natural disasters in 2017: Lower mortality, higher cost.* Retrieved June 26, 2018, from http://cred.be/sites/default/files/CredCrunch50.pdf

European Commission. (2014). *Commission regulation (eu) no 1312/2014.* Retrieved May 03, 2018, from http://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:32014R1312&from=EN

European Commission. (2018). *About INSPIRE.* Available from https://inspire.ec.europa.eu/about-inspire/563.

Galloway, G. E. (2008). Flood risk management in the united states and the impact of hurricane katrina. *International Journal of River Basin Management*, *6*(4), 301-306. doi: 10.1080/15715124.2008.9635357

Gruber, T. R. (1993). A translation approach to portable ontology specifications. *Knowledge Acquisition*, *5*(2), 199–220. doi: 10.1006/knac.1993.1008

GSDI. (2018). *The Future of GSDI.* Available from `http://www.gsdiassociation.org`.

GSDI Technical Working Group. (2009). *Developing Spatial Data Infrastructures: The SDI Cookbook* (D. D. Nebert, Ed.). GSDI-Technical Working Group.

Guha, R. V., Brickley, D., & Macbeth, S. (2016, January). Schema.org: Evolution of structured data on the web. *Communications of the ACM*, *59*(2), 44–51. doi: 10.1145/2844544

Haarslev, V., Hidde, K., Möller, R., & Wessel, M. (2012). The racerpro knowledge representation and reasoning system. *Semantic Web*, *3*(3), 267–277.

Hartmann, T., & Driessen, P. (2017). The flood risk management plan: towards spatial water governance. *Journal of Flood Risk Management*, *10*(2), 145-154. doi: 10.1111/jfr3.12077

Hendriks, P. H., Dessers, E., & van Hootegem, G. (2012). Reconsidering the definition of a spatial data infrastructure. *International Journal of Geographical Information Science*, *26*(8), 1479–1494. doi: 10.1080/13658816.2011.639301

Hu, C., Li, J., Lin, X., Chen, N., & Yang, C. (2018). An observation capability semantic-associated approach to the selection of remote sensing satellite sensors: A case study of flood observations in the jinsha river basin. *Sensors*, *18*(5). Retrieved from `https://doi.org/10.3390/s18051649` doi: 10.3390/s18051649

Hung, H.-C., Lu, Y.-T., & Hung, C.-H. (2018). The determinants of integrating policy-based and community-based adaptation into coastal hazard risk management: a resilience approach. *Journal of Risk Research*, *0*(0), 1-19. doi: 10.1080/13669877.2018.1454496

Jang, M., & Sohn, J.-C. (2004). Bossam: An extended rule engine for owl inferencing. In *International workshop on rules and rule markup languages for the semantic web* (pp. 128–138).

Janowicz, K., Schade, S., Bröring, A., Keßler, C., Maué, P., & Stasch, C. (2010).

Semantic enablement for spatial data infrastructures. *Transactions in GIS*, *14*(2), 111–129. doi: 10.1111/j.1467-9671.2010.01186.x

Kaur, H., Gupta, S., Parkash, S., & Thapa, R. (2018). Application of geospatial technologies for multi-hazard mapping and characterization of associated risk at local scale. *Annals of GIS*, *24*(1), 33-46. doi: 10.1080/19475683.2018.1424739

Kmoch, A., Klug, H., Ritchie, A. B., Schmidt, J., & White, P. A. (2016). A spatial data infrastructure approach for the characterization of new zealand's groundwater systems. *Transactions in GIS*, *20*(4), 626–641.

Kumar, S., & Kumar, S. (2014). Semantic web attacks and countermeasures. In *2014 international conference on advances in engineering technology research (icaetr - 2014)* (p. 1-5). doi: 10.1109/ICAETR.2014.7012841

Lee, C.-S., Kao, Y.-F., Kuo, Y.-H., & Wang, M.-H. (2007). Automated ontology construction for unstructured text documents. *Data & Knowledge Engineering*, *60*(3), 547–566.

Li, J. (2018). *DATA601 report: A web-application for rainfall data analysis.* (University of Canterbury)

Lourdusamy, R., & John, A. (2018). A review on metrics for ontology evaluation. In *2018 2nd international conference on inventive systems and control (icisc)* (p. 1415-1421). doi: 10.1109/ICISC.2018.8399041

Lutz, M., Sprado, J., Klien, E., Schubert, C., & Christ, I. (2009). Overcoming semantic heterogeneity in spatial data infrastructures. *Computers Geosciences*, *35*(4), 739 - 752. (Geoscience Knowledge Representation in Cyberinfrastructure) doi: https://doi.org/10.1016/j.cageo.2007.09.017

Ma, X., Fu, L., West, P., & Fox, P. (2018). Ontology usability scale: Context-aware metrics for the effectiveness, efficiency and satisfaction of ontology uses. *Data Science Journal*, *17*, 10. doi: http://doi.org/10.5334/dsj-2018-010

Martínez-Graña, A., Gómez, D., Santos-Francés, F., Bardají, T., Goy, J. L., & Zazo, C. (2018). Analysis of flood risk due to sea level rise in the menor sea (murcia, spain. *Sustainability*, *10*(3), 780.

Masser, I. (2017). Evaluating the performance of large scale sdis: two contrasting approaches. *International Journal of Spatial Data Infrastructures Research*, *12*, 26–38. doi: 10.2902/1725-0463.2017.12.art2

Metcalfe, P., Beven, K., Hankin, B., & Lamb, R. (2018). A new method, with

application, for analysis of the impacts on flood risk of widely distributed enhanced hillslope storage. *Hydrology and Earth System Sciences*, *22*(4), 2589–2605. Retrieved from https://www.hydrol-earth-syst-sci.net/22/2589/2018/ doi: 10.5194/hess-22-2589-2018

Nadarajah, S., & Choi, D. (2007, Aug 01). Maximum daily rainfall in south korea. *Journal of Earth System Science*, *116*(4), 311–320. Retrieved from https://doi.org/10.1007/s12040-007-0028-0 doi: 10.1007/s12040-007-0028-0

New Zealand Government. (2011). *Spatial Data Infrastructure Cookbook* (Tech. Rep.). Retrieved August 24, 2018, from http://www.gsdiassociation.org/images/publications/cookbooks/SDI_Cookbook_New_Zealand_v1-1_17Nov2011.pdf

Olyazadeh, R., Sudmeier-Rieux, K., Jaboyedoff, M., Derron, M.-H., & Devkota, S. (2017). An offline–online web-gis android application for fast data acquisition of landslide hazard and risk. *Natural Hazards and Earth System Sciences*. doi: 10.5194/nhess-17-549-2017

Perez, J., Arenas, M., & Gutierrez, C. (2006, 06). Semantics and complexity of sparql. In *Acm transactions on database systems* (Vol. 34). doi: 10.1007/11926078_3

Ran, J., & Nedovic-Budic, Z. (2016). Integrating spatial planning and flood risk management: A new conceptual framework for the spatially integrated policy infrastructure. *Computers, Environment and Urban Systems*, *57*, 68 - 79. Retrieved from http://www.sciencedirect.com/science/article/pii/S0198971516300084 doi: https://doi.org/10.1016/j.compenvurbsys.2016.01.008

Rauter, M., & Winkler, D. (2018). Predicting natural hazards with neuronal networks. *arXiv preprint arXiv:1802.07257*.

Razzaq, A., Latif, K., Ahmad, H. F., Hur, A., Anwar, Z., & Bloodsworth, P. C. (2014). Semantic security against web application attacks. *Information Sciences*, *254*, 19 - 38. Retrieved from http://www.sciencedirect.com/science/article/pii/S0020025513005677 doi: https://doi.org/10.1016/j.ins.2013.08.007

Reed, C. (2017). *OGC, INSPIRE, and Metadata* [Blog post]. Retrieved May 03,

2018, from http://www.opengeospatial.org/blog/2630

Röthlisberger, V., Zischg, A. P., & Keiler, M. (2017). Identifying spatial clusters of flood exposure to support decision making in risk management. *Science of The Total Environment*, *598*, 593 - 603. doi: https://doi.org/10.1016/j.scitotenv.2017.03.216

Sayers, P., Penning-Rowsell, E. C., & Horritt, M. (2018, Feb 01). Flood vulnerability, risk, and social disadvantage: current and future patterns in the uk. *Regional Environmental Change*, *18*(2), 339–352. doi: 10.1007/s10113-017-1252-z

Schmachtenberg, M., Bizer, C., & Paulheim, H. (2014). Adoption of the linked data best practices in different topical domains. In *International semantic web conference* (pp. 245–260).

SDW. (2017). *Spatial Data on the Web Best Practices.* Available from https://www.w3.org/TR/sdw-bp/.

Shearer, R., Motik, B., & Horrocks, I. (2008). Hermit: A highly-efficient owl reasoner. In *Owled* (Vol. 432, p. 91).

Simperl, E., & Luczak-Rösch, M. (2014). Collaborative ontology engineering: a survey. *The Knowledge Engineering Review*, *29*(1), 101–131. doi: 10.1017/S0269888913000192

Sirin, E., Parsia, B., Grau, B. C., Kalyanpur, A., & Katz, Y. (2007). Pellet: A practical owl-dl reasoner. *Web Semantics: science, services and agents on the World Wide Web*, *5*(2), 51–53.

Steiniger, S., & Hunter, A. J. (2012). Free and open source gis software for building a spatial data infrastructure. *Geospatial free and open source software in the 21st century*, 247–261.

Stowe, K., Palmer, M., Anderson, J., Kogan, M., Palen, L., Anderson, K. M., ... Lazrus, H. (2018). Developing and evaluating annotation procedures for twitter data during hazard events. In *Proceedings of the joint workshop on linguistic annotation, multiword expressions and constructions (law-mwe-cxg-2018)* (pp. 133–143).

Teng, J., Jakeman, A., Vaze, J., Croke, B., Dutta, D., & Kim, S. (2017, 04). Flood inundation modelling: A review of methods, recent advances and uncertainty analysis. *Environmental Modelling and Software*, *90*, 201-216.

Tsarkov, D., & Horrocks, I. (2006). Fact++ description logic reasoner: System

description. In *International joint conference on automated reasoning* (pp. 292–297).

Vousdoukas, M. I., Voukouvalas, E., Mentaschi, L., Dottori, F., Giardino, A., Bouziotas, D., ... Feyen, L. (2016). Developments in large-scale coastal flood hazard mapping. *Natural Hazards and Earth System Sciences*, *16*(8), 1841–1853.

W3C. (2004). *OWL Web Ontology Language Overview*. Retrieved July 30, 2018, from https://www.w3.org/TR/2004/REC-owl-features-20040210/

W3C. (2012). *OWL 2 web ontology language new features and rationale* (2nd ed.). Retrieved July 30, 2018, from http://www.w3.org/TR/2012/REC-owl2-new-features-20121211

W3C. (2013a). *Linked Data Cookbook*. Available from https://www.w3.org/2011/gld/wiki/Linked_Data_Cookbook.

W3C. (2013b). *Linked Data Glossary*. Available from https://dvcs.w3.org/hg/gld/raw-file/default/glossary/index.html.

W3C. (2013c). *SPARQL 1.1 Overview*. Retrieved July 30, 2018, from https://www.w3.org/TR/2013/REC-sparql11-overview-20130321/

W3C. (2014a). *RDF 1.1 primer*. Retrieved July 30, 2018, from https://www.w3.org/TR/2014/NOTE-rdf11-primer-20140624/

W3C. (2014b). *RDF schema 1.1*. Retrieved July 30, 2018, from https://www.w3.org/TR/2014/REC-rdf-schema-20140225/

W3C. (2017a). *Ontology rights and licence*. Available from https://www.w3.org/2015/spatial/wiki/Ontology_rights_and_licence.

W3C. (2017b). *Shape Expressions (ShEx) Primer*. Retrieved July 30, 2018, from https://shex.io/shex-primer-20170713/

W3C. (2017c). *Shapes Constraint Language (SHACL)*. Retrieved July 30, 2018, from https://www.w3.org/TR/2017/REC-shacl-20170720/

Wächter, T., & Schroeder, M. (2010). Semi-automated ontology generation within obo-edit. *Bioinformatics*, *26*(12), i88–i96.

Wallis, J., Rolando, E., & Borgman, C. (2013). If we share data, will anyone use them? data sharing and reuse in the long tail of science and technol-

ogy. *PLoS ONE*, *8*(7). Retrieved from https://doi.org/10.1371/journal.pone.0067332

Williamson, I. (2003). *Developing spatial data infrastructures:from concept to reality* (I. Williamson, A. Rajabifard, & M.-E. F. Feeney, Eds.). London, New York: Taylor Francis.

Wrachien, D. D., Garrido, J., Mambretti, S., & Requena, I. (2012). Ontology for flood management: a proposal. In *Flood recovery, innovation and response III* (Vol. 159). WIT Press. doi: 10.2495/friar120011

Xiang, Z., Zheng, J., Lin, Y., & He, Y. (2015, Jan 09). Ontorat: automatic generation of new ontology terms, annotations, and axioms based on ontology design patterns. *Journal of Biomedical Semantics*, *6*(1), 4. Retrieved from https://doi.org/10.1186/2041-1480-6-4 doi: 10.1186/2041-1480-6-4

Yu, L., & Liu, Y. (2015). Using linked data in a heterogeneous sensor web: challenges, experiments and lessons learned. *International Journal of Digital Earth*, *8*(1), 17-37. doi: 10.1080/17538947.2013.839007

Appendices

A Relevant Ontologies

While ontologies will need to be tailored to the given use case, it is good practice to reuse existing ontologies as much as possible, so that the same terms and concepts are used at a global level facilitating global interoperability. For this reason, this section lists some relevant ontologies in regards to risk-hazard management.

List of relevant ontologies:

Name: Geo
Purpose: To represent latitudes and longitudes in the WGS84 geodetic reference
URL: `http://www.w3.org/2003/01/geo/wgs84_pos`

Name: DCAT
Purpose: To represent data catalogs
URL: `https://www.w3.org/TR/vocab-dcat/`

Name: DublinCore
Purpose: To represent various types of metadata
URL: `http://dublincore.org/documents/dcmi-terms/`

Name: DQV
Purpose: To annotate quality of data
URL: `https://www.w3.org/TR/vocab-dqv/`

Name: FOAF
Purpose: To represent people entities
URL: `http://xmlns.com/foaf/spec/`

Name: GeoNames
Purpose: To add geospatial information
URL: http://www.geonames.org/ontology/documentation.html

Name: GeoSPARQL
Purpose: To represent spatial objects and their geometries
URL: http://geosparql.org/

Name: GeoSPARQL functions
Purpose: Represents spatial functions in GeoSPARQL
URL: http://www.opengis.net/ont/sf#

Name: GML
Purpose: To represent GML in LD format
URL: http://www.opengis.net/ont/gml#

Name: MOAC
Purpose: To represent entities related to crisis management activities as Linked Data
URL: http://www.observedchange.com/moac/ns/

Name: LOCN
Purpose: To describe places
URL: https://www.w3.org/ns/locn

Name: OWL2
Purpose: High-level expressiveness of LD
URL: https://www.w3.org/TR/owl2-primer/

Name: PROV-O
Purpose: To represent the provenance of LD

URL: https://www.w3.org/TR/prov-o/

Name: RDFS
Purpose: Basic expression of LD
URL: https://www.w3.org/TR/rdf-schema/

Name: SKOS
Purpose: To represent hierarchies of data
URL: https://www.w3.org/TR/skos-primer/

Name: Time
Purpose: To represent time
URL: https://www.w3.org/TR/owl-time/

Name: Vcard
Purpose: To describe people and organizations
URL: http://www.w3.org/2006/vcard/ns#

Name: VoID
Purpose: To represent LD metadata
URL: https://www.w3.org/TR/void/

B Linked Data Stores: Review and Recommendation

Linked Data stores are needed to effectively store and query RDF data. However, the linked data stores should also allow for spatial querying for this project's use case. In this section, existing linked data stores are reviewed based on their advertised features, and various user reviews found on the Web.

AllegroGraph

AllegroGraph[27] is a triple store that is used in both open source and commercial projects. It has both a commercial and a free but limited version, and can be used with Java, Python, Ruby, Perl, C#, Clojure, and Common Lisp. The main feature of AllegroGraph is its ability to scale to 'billions of quads while maintaining superior performance' [28]. There is also an Amazon Machine Image (AMI) allowing for the easy use of AllegroGraph using Amazon's server. It supports SPIN, which allows functions to be called in a SPARQL query. This feature makes any sort of queries very powerful as they can be highly customised. AllegroGraph also offers RDFS reasoning supporting all of RDF and RDFS predicates and selected ones from OWL. In AllegroGraph, both static and dynamic materialisation of data inferences can be used. In addition, Prolog is also included, which would allow high-level rules to be expressed.

However, to make use of AllegroGraph's most powerful features such as federation, warm standby, Point in Time recovery, Replication, and triples level security, it needs to be bought. The free version limits the triples to 5 Million, the developer tier limits it to 50 Million, and the enterprise level has unlimited number of triples.

[27] https://allegrograph.com/
[28] https://franz.com/agraph/allegrograph/

BlazeGraph

BlazeGraph is an open source graph database that supports RDF and SPARQL. It is the first GPU-accelerated database for large graphs. It does not support GeoSPARQL, but can store spatial objects and have the basic capabilities for simple spatial queries.

GraphDB

GraphDB comes with different versions – Free, standard, and enterprise. The free version has full capabilities except that it can only handle two queries in parallel at a time. It supports all the features of a triple store, such as querying, inferencing, and reasoning. It performs querying and reasoning using file-based indices, meaning that the data are grouped by documents and those documents are indexed for fast processing. It is also available as a docker image. However, GraphDB only supports within and nearby spatial queries.

Jena Triple Database (TDB)

Apache Jena is an open source project for the Semantic Web. It supports RDF, reasoning and inferences. Jena TDB focuses on quick data access, not insertion. Open source projects provide more freedom as it allows total control of the code. TDB supports spatial indexing, as well as, GeoSPARQL, but the functions are currently limited to: east, west, north, south, intersectsBox, isWithinBox, isWithinCircle, and isNearBy [29]. Further, these spatial operators cannot compare different shapes, only shapes to literal values.

Neo4J

Neo4J is a free pure-graph database. As RDF data models are directed labelled graphs, using a graph database can be utilised if features pertaining to graphs

[29] `https://jena.apache.org/documentation/query/spatial-query.html`

are required (e.g. fast traversal, nearest neighbour, degrees of separation, shortest path algorithms). While Neo4J can be extended to reason over RDF data, reasoning which is an important part of RDF does not come natively, neither does SPARQL—a plug-in was developed though. Nonetheless, Neo4J can support these spatial operators: contain, cover, covered by, cross, disjoint, intersect, intersect windows, overlap, touch, within, and within distance.

RDF4J Core Databases

This database is developed by the RDF4J group[30], and is advertised as being intended for small to medium-sized databases. The amount of triples it can efficiently process is up to 100 Million. RDF4J itself is a framework to work with RDF data, and as such RDf4J databases are compatible with their framework. The database is free, and supports the most spatial operators in this list such as: boundary, ocnvexHull, difference, ehContains, ehCoveredBy, ehCovers, ehCoveredBy, ehCovers, ehDisjoint, ehInside, ehMeet, ehOverlap, envelope, equals, intersection, relate, sfContains, sfCrosses, sfDisjoint, sfIntersects, sfOverlaps, sfTouches, sfWithin, symDifference, and union. Further, it includes reasoning of SHACL and SPIN.

Stardog

Stardog[31] is a commercial product that is easy to set-up and use. Their base is a graph database that can reason, infer, do spatial queries, and search text semantically. However the spatial features are not part of the free version, and are currently limited to relate, distance, within, nearby, and area operators.

[30] `http://rdf4j.org/`
[31] `https://www.stardog.com/`

Parliament

Parliament[32] is an open-source triple store (under the BSD license) designed for the Semantic Web. It boasts of efficient querying and insertion of data by reordering query execution to start with the most restrictive query first. In addition, Parliament supports GeoSPARQL and temporal queries. The most important aspect of an ontology for automation is the inferences that can be derived from existing data. Parliament makes use of a rule engine as a means for fast inferences of new data. Their rule engine implements RDFS inferences and some selected elements of OWL. However, it is also possible to add further man-made rules for more specific inferences. Being an RDF database, Parliament provides a SPARQL endpoint that is decoupled from the server. The spatial operators it supports are: contains, crosses, disjoint, equals, intersects, overlaps, touches, and within.

Virtuoso

Virtuoso is presented as 'a modern enterprise-grade solution for data access, virtualization, integration and multi-model relational database management (SQL Tables and/or RDF Statement Graphs)' `https://virtuoso.openlinksw.com/`. Virtuoso is composed of multiple engines such as RDBMS, virtual databases, messaging and storage protocols, and reasoning and inferences. The virtual database engine acts as a middle-man to access multiple other databases and hence, allowing join operations across federated databases of multiple types. It has both open source and commercial licenses by OpenLink Software. It uses an SQL Relational Database Management System (RDBMS) as its core. Available spatial functions in Virtuoso are: intersects, within, bounding box, as text, and distance.

[32] `http://parliament.semwebcentral.org/`

Recommendation

While there are a lot of linked data stores available, the important factors would be the ease of use and flexibility of the chosen solution. The database should be flexible enough to allow the developers to inject code that is not currently available, but it also needs to be a fully functional triple store with querying, reasoning, and inferencing included natively. In addition, the familiarity of a programming language is important for its ease of adoption alongside the cost of the program.

For these reasons, the recommended database should be free to use and modify (i.e. open source), and be in a common programming language. It must support spatial queries to a reasonable extent, and have good documentations. For these reasons, RDF4J is recommended. RDF4J uses Java, a common programming language, and it supports numerous GeoSPARQL operators, alongside reasoners for SHACL (W3C, 2017c) and further has its own framework. While virtuoso, Stardog, and AllegroGraph have promising features, they are commercial by design, and hence not recommended to prevent a 'vendor lock-in'. Parliament is not recommended because it has not been updated since 2015, and hence might be outdated and not maintained any more. Jena Triple Database could be a good option, however, its supported spatial operators are lacking. As for BlazeGraph, it does not support spatial queries and hence is not recommended.

However, RDF4J databases are designed for small to medium-sized databases for up to 100 million triples. If more triples need to be stored, commercial software might need to be purchased, and further reviews will need to be conducted. For the purpose of a prototype, RDF4J should suffice.